

# Comparability *of* Large-Scale Educational Assessments

ISSUES *and* RECOMMENDATIONS



NATIONAL ACADEMY *of* EDUCATION

# **Comparability *of* Large-Scale Educational Assessments**

---

ISSUES *and* RECOMMENDATIONS

Amy I. Berman, National Academy of Education

Edward H. Haertel, Stanford University

James W. Pellegrino, University of Illinois at Chicago

National Academy of Education  
Washington, DC

NATIONAL ACADEMY OF EDUCATION 500 Fifth Street, NW

Washington, DC 20001

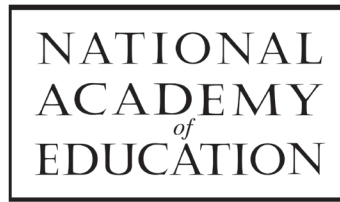
NOTICE: This project and research reported here were supported by a grant from Smarter Balanced/University of California, Santa Cruz. The opinions expressed are those of the editors and authors and do not represent the views of Smarter Balanced/University of California, Santa Cruz.

Digital Object Identifier: 10.31094/2020/1

Additional copies of this publication are available from the National Academy of Education, 500 Fifth Street, NW, Washington, DC 20001; <http://www.naeducation.org>.

Copyright 2020 by the National Academy of Education. All rights reserved.

Suggested citation: Berman, A. I., Haertel, E. H., & Pellegrino, J. W. (2020). *Comparability of Large-Scale Educational Assessments: Issues and Recommendations*. Washington, DC: National Academy of Education.



The **National Academy of Education** (NAEd) advances high-quality research to improve education policy and practice. Founded in 1965, the NAEd consists of U.S. members and international associates who are elected on the basis of scholarship related to education. The Academy undertakes research studies to address pressing educational issues and administers professional development fellowship programs to enhance the preparation of the next generation of education scholars.



## COMPARABILITY OF LARGE-SCALE EDUCATIONAL ASSESSMENTS: ISSUES AND RECOMMENDATIONS

### *Steering Committee*

**Edward H. Haertel** (*Co-Chair*), Graduate School of Education, Stanford University  
**James W. Pellegrino** (*Co-Chair*), Learning Sciences Research Institute, University of  
Illinois at Chicago

**Louis M. Gomez**, Graduate School of Education and Information Studies, University  
of California, Los Angeles

**Larry V. Hedges**, Department of Statistics, Northwestern University

**Joan L. Herman**, National Center for Research on Evaluation, Standards, and  
Student Testing, University of California, Los Angeles

**Diana C. Pullin**, Lynch School of Education and Human Development, Boston  
College

**Marshall S. Smith**, Carnegie Foundation for the Advancement of Teaching

**Guadalupe Valdes**, Graduate School of Education, Stanford University

### *Staff*

**Amy I. Berman**, Deputy Director



# Contents

<b>EXECUTIVE SUMMARY</b>	<b>1</b>
<b>1 INTRODUCTION: FRAMING THE ISSUES</b>	<b>9</b>
<i>Amy Berman, National Academy of Education; Edward Haertel, Stanford University; and James Pellegrino, University of Illinois at Chicago</i>	
<b>2 COMPARABILITY OF INDIVIDUAL STUDENTS' SCORES ON THE "SAME TEST"</b>	<b>25</b>
<i>Charles DePascale and Brian Gong, National Center for the Improvement of Educational Assessment (Center for Assessment)</i>	
<b>3 COMPARABILITY OF AGGREGATED GROUP SCORES ON THE "SAME TEST"</b>	<b>49</b>
<i>Leslie Keng and Scott Marion, Center for Assessment</i>	
<b>4 COMPARABILITY WITHIN A SINGLE ASSESSMENT SYSTEM</b>	<b>75</b>
<i>Mark Wilson, University of California, Berkeley, and Richard Wolfe, Ontario Institute for Studies in Education, University of Toronto</i>	
<b>5 COMPARABILITY ACROSS DIFFERENT ASSESSMENT SYSTEMS</b>	<b>123</b>
<i>Marianne Perie, Measurement in Practice, LLC</i>	
<b>6 COMPARABILITY WHEN ASSESSING ENGLISH LEARNER STUDENTS</b>	<b>149</b>
<i>Molly Faulkner-Bond, WestEd, and James Soland, University of Virginia/Northwest Evaluation Association (NWEA)</i>	
<b>7 COMPARABILITY WHEN ASSESSING INDIVIDUALS WITH DISABILITIES</b>	<b>177</b>
<i>Stephen Sireci and Maura O'Riordan, University of Massachusetts Amherst</i>	
<b>8 COMPARABILITY IN MULTILINGUAL AND MULTICULTURAL ASSESSMENT CONTEXTS</b>	<b>205</b>
<i>Kadriye Ercikan, Educational Testing Service/University of British Columbia, and Han-Hui Por, Educational Testing Service</i>	
<b>9 INTERPRETING TEST-SCORE COMPARISONS</b>	<b>227</b>
<i>Randy Bennett, Educational Testing Service</i>	
<b>BIOGRAPHICAL SKETCHES OF STEERING COMMITTEE MEMBERS AND AUTHORS</b>	<b>237</b>





# Executive Summary

“How is my child doing?” “What are my child’s strongest and weakest subjects?”  
“Have my child’s test scores improved from last year?” “How does my child’s  
test scores compare to others looking to go to college?” “Should I move to this  
school zone?”

—Parent questions

“How do the assessment scores of schools within our district compare?” “How  
are our English learner students doing compared with our native English speak-  
ers?” “Are we closing the achievement gap?” “How do our assessment scores  
compare to others within the state?”

—District administrator questions

“How do our kids measure up to kids in other states?” “Within districts?” “How  
are the scores of various student subgroups changing over time?”

—State administrator/policy maker questions

## PURPOSE OF THE VOLUME

Such questions come from a range of stakeholders with separate vested interests in educational assessments, ranging from parents worried about individual test scores, to local district leaders interested in specific populations, to state policy makers looking at the big picture. Often, different questions are asked about the same assessments, and these questions do not always coincide with the uses for which the assessments were designed and validated. While their interests and questions may differ, these

stakeholders all have one thing in common: they are asking questions that assume scores can be validly compared—that a lower score means less proficiency, similar scores mean similar proficiency, a higher score means greater proficiency, and a positive change in scores from one year to the next means improvement, regardless of the specific details of how each student was tested. In other words, they *assume* the *comparability* of scores from educational assessments.<sup>1</sup>

Stakeholders often simply assume that scores obtained from different students in different times and places, using different tests or test forms, are directly comparable, but that is not necessarily the case. Countless factors influence assessment scores, and finding accurate and satisfactory answers to questions of score comparability is not easy. Moreover, comparability may be adequate for one interpretive purpose but not for another.

This National Academy of Education (NAEd) volume provides guidance to key stakeholders on how to accurately report and interpret comparability assertions as well as how to ensure greater comparability by paying close attention to key aspects of assessment design, content, and procedures. The goal of the volume is to provide guidance to relevant state-level educational assessment and accountability decision makers, leaders, and coordinators; consortia members; technical advisors; vendors; and the educational measurement community regarding *how much* and *what types* of variation in assessment content and procedures can be allowed, while still maintaining comparability across jurisdictions and student populations. At the same time, the larger takeaways from this volume will hopefully provide guidance to policy makers using assessment data to enact legislation and regulations and to district- and school-level leadership to determine resource allocations, and also provide greater contextual understanding for those in the media using test scores to make comparability determinations.

## WHAT IS COMPARABILITY?

Users of educational tests often seek to compare scores even if the scores were obtained at different times, in different places, or using variations in assessment content and procedures. Score comparability broadly means that users can be confident in making such comparisons. Ideally, users could be assured that students with the same score are equally proficient with respect to the knowledge and skills a test was intended to measure. As described more fully throughout this volume, there are numerous threats to comparability that must be considered before making such a claim. For instance, if test performance requires proficiencies irrelevant to the knowledge and skills the test is intended to measure, and if some students' performance suffers due to lack of those proficiencies, then the scores of those students are not comparable to the scores of other students (e.g., a math test may not be intended to test language proficiency, but limited language knowledge may nonetheless influence test results for some students). Threats to comparability may also arise due to differences in test administration or scoring conditions (e.g., paper-and-pencil versus computer-based testing, different times in the academic year, or human versus machine scoring), and differences in the specific

---

<sup>1</sup> The words *assessment* and *test* are used throughout this volume, and though to some extent they are interchangeable, they do have different meanings. *Assessment* is the more general of the words, conveying the idea of a process providing evidence of quality. *Assessment* covers a broad range of procedures to measure teaching and learning. A *test* is one product that measures a particular set of objectives or behavior.

test or test form used. When scores are compared for groups of students, comparability also demands that the groups compared be defined consistently, with proper attention to sampling, rules for exclusions and exemptions, and retesting practices. Finally, the issue of score comparability requires attention to the inferences drawn from test scores, as well as the intended uses of the tests. What does the end user want to compare, at what aggregated level, and for what purpose?

### WHAT IS IN THE VOLUME?

The volume is organized by the major types of comparisons that end users often examine—comparability of individual students' scores, of aggregated group scores, within a single assessment system, and across different assessment systems. While issues specific to certain groups of students exist within each of the major types of comparisons (and are addressed across the chapters), the volume also includes chapters dedicated to examining comparability issues specific to certain subgroups and populations, including English learner (EL) students, students with disabilities, and populations encompassing students with differing linguistic and cultural backgrounds.

Chapter	Title	Authors
1	Introduction: Framing the Issues	Amy Berman, National Academy of Education; Edward Haertel, Stanford University; and James Pellegrino, University of Illinois at Chicago
2	Comparability of Individual Students' Scores on the "Same Test"	Charles DePascale and Brian Gong, National Center for the Improvement of Educational Assessment (Center for Assessment)
3	Comparability of Aggregated Group Scores on the "Same Test"	Leslie Keng and Scott Marion, Center for Assessment
4	Comparability Within a Single Assessment System	Mark Wilson, University of California, Berkeley, and Richard Wolfe, Ontario Institute for Studies in Education
5	Comparability Across Different Assessment Systems	Marianne Perie, Measurement in Practice
6	Comparability When Assessing English Learner Students	Molly Faulkner-Bond, WestEd, and James Soland, University of Virginia/Northwest Evaluation Association (NWEA)
7	Comparability When Assessing Individuals with Disabilities	Stephen Sireci and Maura O'Riordan, University of Massachusetts Amherst
8	Comparability in Multilingual and Multicultural Assessment Contexts	Kadriye Ercikan, Educational Testing Service/University of British Columbia, and Han-Hui Por, Educational Testing Service
9	Interpreting Test-Score Comparisons	Randy Bennett, Educational Testing Service

## MAJOR FINDINGS FROM THIS VOLUME

Comparability often operates on the assumption that all students' scores come from the same test; however, whether such thinking is focused on results within one school, across districts, across states, or across time, scores from the "same" test may be influenced by many factors, including different item pools, timing of the administration (a few weeks' difference can make a big difference in students' opportunity to learn), test administration conditions, and accommodations. It may be unclear what is or is not the "same" test. Moreover, intended comparisons often span scores from entirely different tests. Whether these comparisons are across states using different statewide assessments or across different countries using similar but adapted tests, comparability is compromised. In general, comparisons are most defensible when the same assessment is given under substantively the same conditions to similar student samples at the same point in time. The legitimacy of comparisons thus becomes less certain as the assessment, the assessment conditions, student samples, and the time of administration diverge. Thus, there is a continuum where comparisons for certain purposes are appropriate and reasonable and some comparisons should not be made. Throughout this volume is a wealth of examples across this continuum.

This volume offers many recommendations to help improve comparability and inform judgments about comparability claims across jurisdictions, among populations, and over time. Here we highlight some of the major findings across the chapters, organized by these cross-cutting themes: (1) Purpose, Design, and Interpretation; (2) Content and Construct Domain; (3) Measurement Properties; (4) Administration Conditions; and (5) Student Background Factors: Experiential, Linguistic, and Sociocultural.<sup>2</sup>

### Purpose, Design, and Interpretation

*Clearly define and communicate the intended purposes and uses of tests.* Comparability and validity are contingent on the intended purposes and uses of test scores, as well as the score-based claims users wish to make. Moreover, students' motivations, and therefore scores, will vary based on the purpose of a test (e.g., students may be more motivated when taking a test used in college admission decisions than for the National Assessment of Educational Progress (NAEP), which is a nationally sampled test for which students never receive a score). It is therefore essential to communicate with the field and end users about appropriate and inappropriate score interpretations.

*Be explicit about the design of the assessment and assessment system.* From the beginning, the design of the assessment system should reflect its purposes and intended interpretive uses. Where possible, the design should anticipate and attempt to mitigate unintended consequences, including those that may arise

---

<sup>2</sup> In 2014, the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) issued the most current version of guidelines and standards for educational and psychological testing, many of which elaborate on points discussed here. AERA, APA, & NCME. (2014). *Standards for educational and psychological measurement*. Washington, DC: AERA.

from unintended but foreseeable score uses. Ideally, developers should address potential threats to comparability during the development of the assessment system; evaluate the degree to which the threats are mitigated as the system is implemented; and then, if necessary, respond appropriately if comparability proves inadequate to support widespread or important uses.

***Evaluate comparability for each intended interpretation.*** There may be variations in test items, administration conditions, scoring procedures, students' opportunity to learn (OTL), students' familiarity with the test itself, and/or the intended uses of the test results. These variations, individually and cumulatively, influence student performance and can affect score comparability. Comparability may be adequate for one use or purpose but not for another. Determining whether there is sufficient test-score comparability to support a given score interpretation involves consideration of the technical methods used to account for any such variations, as well as the rationale for the interpretation, the logic and evidence that support it (including identifying divergences and justifying the rationale for making score comparisons in light of the divergences), and the consequences (stakes) associated with the interpretation. Test developers and assessment system designers should provide users and potential audiences with clear guidance for appropriate and inappropriate score uses and inferences.

***Ensure groups are properly defined.*** When groups are compared, group definitions and sampling procedures should be precise so that comparisons are not distorted by irrelevant background factors. If rules for testing exclusions or exemptions, or retesting practices, differ or are differently implemented from place to place, then group comparisons may be distorted.

### Content and Construct Domain

***Clearly define the subject-matter content of an assessment and an assessment system.*** A clear content framework or blueprint for the assessment and assessment system is essential in defining the basis and objective for comparability. It provides, in detail, the content, content breakdowns and specifications, assessment goals and priorities, item types, and numbers of items per content category. It should also align with instruction. The clearer the blueprint, the easier it is to look across tests to evaluate comparability for any given use or interpretation.

***Examine what it is that tests actually measure (i.e., constructs) when making comparability claims.*** To make comparability claims, it is important to examine not only the content of the assessments (i.e., that the tests measure the appropriate content), but also that the items validly measure the intended skills and/or abilities and at similar levels of depth or challenge. The constructs tapped by a set of items depend not only on item content, but also on what the respondent must do with that content—one item may require only rote recall while another calls for complex reasoning. Note also that users of test scores may hold assumptions about what test scores mean, which are sometimes not fully warranted.

Claims that stay close to the content and processes actually measured will be more defensible than sweeping claims that extrapolate far beyond what the test actually requires students to know and do.

### Measurement Properties

*Stringency is important when comparing across tests, among subject areas, and among grades.* Stringency addresses both the degree of challenge that items pose for test takers and the decision rules for interpreting various levels of test performance. The same set of items may be easy for students at one grade level but difficult for students at a lower grade level. Also, on any test, setting a more stringent “proficiency” definition (i.e., requiring a higher score to reach “proficient”) will yield a lower proportion of students designated “proficient.” Thus, when comparing performance in the same subject matter (e.g., math) across grades, both absolute item difficulties and proficiency levels must be chosen in such a way that the “proficient” proportions do not vary capriciously from one grade to another. Similarly, different stringency levels across subject areas may create an unwarranted perception that achievement in one subject area is lagging.

*The more similar the psychometric characteristics of assessments, the easier it is to have valid linking of tests to support comparability statements.* Often, and with good cause, scores across assessments and assessment systems are compared. For instance, policy makers may want to compare achievement results across states that use different exams. Comparability across assessments is most defensible when the individual tests have similar, high levels of reliability and are designed from the same blueprint to measure the same construct. When this is not true, adequate comparability may still be attained if the tests are designed to provide evidence about the same construct, use the same scaling model, and use similar item types, and if actual score levels are available, as opposed to performance relative to benchmark levels. If groups are to be compared, they should be defined similarly across tests or jurisdictions. The further two or more tests depart from these ideals, the weaker comparability becomes.

### Administration Conditions

*The modes of administration of assessments can affect test results.* Many variables are involved in specifying test administration conditions. A test may be timed or untimed; it may be administered by paper and pencil or by computer. Computer administrations may employ different devices or rely on item selection algorithms devised by different vendors. Various accommodations may be permitted, according to varying criteria. Testing windows may differ relative to school calendars. These and other differences may affect comparability. The more similar the testing conditions, the fewer risks to comparability.

*Accommodations, when appropriately implemented, enhance comparability.* Accommodations provided to students with disabilities and/or EL students are



meant to minimize construct-irrelevant variance (i.e., variance in test scores due to factors other than proficiency with respect to the intended construct). Ideally, an accommodation would function in such a way that, if it were provided to all students, it would improve the scores of students with disabilities who needed it without affecting the scores of other students (interaction hypothesis). If this ideal cannot be attained, then the accommodation should at least improve the scores of those needing it more than it improves the scores of other students (differential boost hypothesis). When properly designed and used, accommodations promote fairness by helping to ensure that the test measures the same intended construct for all students.

***Language that is hard for some students to understand can hinder comparability on tests not intended to measure language skills.*** Such concerns should be explicitly attended to in the assessment design and conditions of administration. For example, EL students, by definition, are still developing their proficiency in English. Thus, it is important to provide accommodations designed to help them demonstrate their construct-relevant knowledge and skills without being hampered by construct-irrelevant limitations in language skills. Recognizing that available accommodations are imperfect, it is important as well to ensure that any comparability statements acknowledge the potential for construct-irrelevant language variance in test scores. To date, language glossaries, particularly when offered with extra testing time, are the only accommodations demonstrated to be effective for EL students without compromising the validity of their responses.

### **Student Background Factors: Experiential, Linguistic, and Sociocultural**

***Ensure familiarity with testing conditions and formats.*** Testing conditions and test formats can compromise the comparability of assessments. For instance, before taking a test using a computer mouse and screen prompts, students should experience and be familiar with these conditions. Ideally, students unfamiliar with timed examinations should be afforded similarly timed experiences prior to taking the test. Students should also have had opportunities to become familiar with item formats prior to encountering them on examinations. Unfamiliar testing conditions and formats threaten comparability. These threats differ from the effects of differential OTL, described below.

***Equivalent opportunity to learn is needed to ensure comparability or, if it differs, to inform comparability statements.*** OTL, with respect to assessments, has been conceptualized as the opportunity to learn what is tested. It includes, among other factors, the consideration of school resources, access to the curriculum, time allocated for instruction, quality of instruction, access to culturally responsive teaching and curriculum and school culture, and students' preparedness to participate in learning. With regard to large-scale assessments, an additional OTL factor pertains to the scheduling of the assessment with respect to completion of instruction. When OTL is not similar, it should be acknowledged



as a threat to comparability. However, if a test is being used purely to describe students' current levels of achievement in the content area, then their scores—regardless of variations in OTL—may support some intended interpretations.

*Students with different linguistic or sociocultural backgrounds should have the same opportunities to demonstrate their knowledge, skills, and competencies on assessments.* When testing programs span diverse language or sociocultural groups, translated versions of tests may be used. However, comparability across translated versions is far from ensured. Items often function differently between language groups, both within and across countries. Even in the same language context, such as in the United States, students from different sociocultural groups may speak structurally and semantically different varieties of the “same” language (e.g., indigenous students, African American students, Mexican American students, and students from nonmainstream socioeconomic backgrounds). The goal with translated tests must be measurement equivalence, including equivalence of construct, test, and testing conditions. The quality of adaptation to other languages is optimized when the assessments in the source language are developed with test adaptation goals in mind.

The chapters to follow delve into these and related issues in greater depth. It is the hope of the authors and editors, and of the NAEEd, that this volume helps to guide wiser and fairer testing policy and practice in education, and in other fields as well.

## Introduction: Framing the Issues

Amy I. Berman, *National Academy of Education*  
 Edward H. Haertel, *Stanford University*  
 James W. Pellegrino, *University of Illinois at Chicago*

“How is my child doing?” “What are my child’s strongest and weakest subjects?”  
 “Have my child’s test scores improved from last year?” “How does my child’s  
 test scores compare to others looking to go to college?” “Should I move to this  
 school zone?”

—Parent questions

“How do the assessment scores of schools within our district compare?” “How  
 are our English learner students doing compared with our native English speak-  
 ers?” “Are we closing the achievement gap?” “How do our assessment scores  
 compare to others within the state?”

—District administrator questions

“How do our kids measure up to kids in other states?” “Within districts?” “How  
 are the scores of various student subgroups changing over time?”

—State administrator/policy maker questions

While such questions are common, finding accurate and satisfactory answers is not an easy task given the countless factors influencing assessment scores. Stakeholders often simply assume that scores obtained from different students in different times and places, using different test forms, are directly comparable. Moreover, the questions come from a range of stakeholders each with a separate vested interest in educational assessments, ranging from parents worried about individual student test scores, to local district leaders interested in a specific population, to state policy makers looking at the broad aggregate data. They are often asking different questions about the same

assessments, but answers to these questions do not always coincide with the interpretive uses for which the assessments were originally designed and validated. While their interests and questions may differ, these stakeholders all have one thing in common: they are asking questions that assume scores can be validly compared—that a lower score means less proficiency, similar scores mean similar proficiency, a higher score means greater proficiency, and a positive change in scores from one year to the next means improvement, regardless of the specific details of how each student was tested. In other words, they *assume* the *comparability* of scores from educational assessments.<sup>1</sup>

And while much of educational news reporting relies on testing data, it is often reported at a high aggregate level without descriptions of the assessments, their purposes, and possible explanatory variables. Recent headlines such as these have the potential to influence people, even if they do not tell the full story: “Minnesota Report Card: Small Schools Score Higher” (Sethrie, 2020); “Maryland’s PARCC Results Show Dip in Math, Improvements in English” (Ryan, 2019); “Oregon Dips in Standardized Test Scores, Mixed Bag for Mid-Valley” (Rimel, 2019); “New Statewide Test Results Show Achievement Gap Throughout Cedar Rapids Community School District” (Kalk, 2019); “Survey: 45% of Test-Takers Boycott ELA Exam [Long Island, NY]” (Tyrrell, 2019); and “Majority of South Bend Schools Do Not Meet Federal Expectations, New Report States” (Kirkman, 2020). Such articles can influence individual decisions concerning where to live or whether to apply to a nontraditional public school as well as state and federal policy makers’ decisions about investments and policies related to educational reform.

This National Academy of Education (NAEd) volume provides guidance to key stakeholders on how to accurately report and interpret comparability assertions as well as how to ensure greater comparability by paying close attention to key aspects of assessment design, content, and procedures. The goal of the volume is to provide guidance to relevant state-level educational assessment and accountability decision makers, leaders, and coordinators; consortia members; technical advisors; vendors; and the educational measurement community regarding *how much* and *what types* of variation in assessment content and procedures can be allowed, while still maintaining comparability across jurisdictions and student populations. At the same time, the larger takeaways from this volume will hopefully provide guidance to policy makers using assessment data to enact legislation and regulations and to district- and school-level leadership to determine resource allocations, and also to provide greater contextual understanding for those in the media using test scores to make comparability determinations.

To accomplish these ambitious goals, the NAEd organized a steering committee comprised of Edward Haertel (Co-Chair), James Pellegrino (Co-Chair), Louis Gomez, Larry Hedges, Joan Herman, Diana Pullin, Marshall S. Smith, and Guadalupe Valdes. The topical foci of the eight chapters following this introduction are the result of the committee’s extensive efforts to determine the most pressing comparability issues currently affecting educational assessment while also ensuring that particular subgroups for which comparability issues often arise are included in the discussion instead of shelved with an asterisk for later discussion. The committee organized these issues into

---

<sup>1</sup> The words *assessment* and *test* are used throughout this volume, and though to some extent they are interchangeable, they do have different meanings. *Assessment* is the more general of the words, conveying the idea of a process providing evidence of quality. *Assessment* covers a broad range of procedures to measure teaching and learning. A *test* is one product that measures a particular set of objectives or behavior.

the following chapters: (1) comparability of individual students' scores on the "same test," (2) comparability of aggregated group scores on the "same test," (3) comparability within a single assessment system, (4) comparability across different assessment systems, (5) comparability when assessing English learner (EL) students, (6) comparability when assessing students with disabilities, (7) comparability in multilingual and multicultural assessment contexts, and (8) interpreting test-score comparisons. The first four chapters progress from narrower to broader interpretive contexts, with comparability claims in each chapter building on those preceding. Chapters 6 through 8 address specific populations meriting additional attention. The final chapter offers a synthesis of best practices for interpreting test-score comparisons. After identifying the chapter themes, the steering committee outlined the chapter goals and identified experts to develop and author the individual chapters. The steering committee, as well as other chapter authors, provided critical feedback on draft chapters, including at a 2-day workshop of authors and the steering committee in June 2019.<sup>2</sup> The results of these efforts comprise this volume.

## BACKGROUND TO THIS VOLUME

Student testing has played an important role in the American education system since its creation. Each day students take tests, most of which are devised by teachers, to monitor student learning and guide instruction. Testing students for the purposes of classroom feedback, system monitoring, and selection and placement decisions have existed for more than 180 years. Standardized written exams began in the mid-19th century (OTA, 1992).

The mid-19th to the mid-20th century served as a time of great expansion for educational testing. Entire books and articles have been written about the history of educational testing (see, e.g., Kaestle, 1983, 2012; OTA, 1992; Resnick, 1982; Vinovskis, 2019). While we cannot do justice to such a history in so short an introduction, we point out that, with both population growth and urbanization, public school enrollment more than doubled from 1870 to 1900 and with it the desire to use educational testing for accountability and classification purposes (OTA, 1992). By 1900, intelligence testing had begun and, following the extensive use of intelligence tests in the Army during World War I, these tests proliferated into American schools (Kaestle, 2012). During the 1920s and 1930s, cost-effective, multiple-choice standardized tests became entrenched in schools (OTA, 1992). And, in 1950, the automatic scoring machine was invented by the Iowa Testing Program and large-scale state and national testing became feasible (OTA, 1992).

Of course we would be remiss in not acknowledging the equity concerns that have abounded in standardized testing. Issues have been raised about the equity (bias) of tests, as well as disparate educational outcomes that result from the use of results from educational tests. Moreover, there is increased diversity of test takers with our ever-changing population as well as expansion of test taking, including greater racial and ethnic diversity, language and cultural diversity, and the inclusion of students with disabilities. While we again cannot do justice to this history in this introduction, others

---

<sup>2</sup> The steering committee also called on the expertise of Christian Faltis to serve as both a discussant at the June 2019 workshop and a reviewer of several chapters. The committee is grateful for his contributions to this volume.

have described these issues (e.g., Moss, Pullin, Gee, Haertel, & Young, 2008; Symposium, 1994) and some of these concerns are raised throughout this volume, including in our chapters addressing English learner students, students with disabilities, and nondominant language and cultural groups.

The Elementary and Secondary Education Act (ESEA) of 1965 included test-based evaluation measures—albeit weak and weakly enforced (Kaestle, 2012; Vinovskis, 2019)—as part of an effort to raise educational achievement and make education more equitable. Then, in 1969, the first national assessments of academic achievement, now known as the National Assessment of Educational Progress (NAEP), were administered.

In 1983, President Ronald Reagan's National Commission on Excellence in Education released *A Nation at Risk*, which asserted that "the educational foundations of our society are presently being eroded by a rising tide of mediocrity that threatens our very future as a Nation and a people," and "[i]f an unfriendly foreign power had attempted to impose on America the mediocre educational performance that exists today, we might well have viewed it as an act of war." This rallying cry led to reform efforts to set high standards and increase accountability measures, often described in the form of testing. In 1991, President Bush proposed the "America 2000" program (and implemented portions through executive order), which called for challenging national standards and voluntary national tests (Vinovskis, 2019). And in 1994, President Clinton's Goals 2000 Act and the Improving America's Schools Act (the latter being the reauthorization of the ESEA) both passed, calling for high educational standards and systems of testing accountability (NAEd, 2009). Finally, in 2002, the federal government mandated annual educational testing in grades 3 through 8 and once in high school for accountability purposes with the passage of the No Child Left Behind Act (NCLB). While NCLB set the impossible goal of all students reaching proficiency on state reading and math tests by 2014, the states' response to NCLB also highlighted the lack of comparability of state standards and assessments.

In 2009, nearly all states, along with the District of Columbia, came together to develop common academic standards in mathematics and English: the Common Core State Standards (CCSS) Initiative. Common standards led to the call for common assessments and, in 2009, through the Race to the Top (RTT) program, the Obama administration announced a competition for grant funding of \$350 million for the development of tests aligned with the CCSS (Jochim & McGuinn, 2016). In 2010, the U.S. Department of Education awarded grants to two state consortia, the Partnership for Assessment of Readiness for College and Careers (PARCC) and the Smarter Balanced Assessment Consortium (Smarter Balanced), which represented 44 states and the District of Columbia, to develop assessment tools aligned with the common standards adopted by states (DOEd, 2010, n.d.; Robelen, 2010).<sup>3</sup> As noted by the U.S. Department of Education in its award letters to PARCC and Smarter Balanced, for public schools to succeed we need "a first-rate assessment system to measure progress, guide instruction, and prepare students for college and careers."<sup>4</sup> Moreover, through RTT, the Obama

<sup>3</sup> There were also several smaller awards to consortia addressing assessments for students with severe disabilities and for English learner students.

<sup>4</sup> The U.S. Department of Education award letters can be found here: <https://www2.ed.gov/programs/racetothetop-assessment/parcc-award-letter.pdf> and <https://www2.ed.gov/programs/racetothetop-assessment/sbac-award-letter.pdf>.

administration offered states a competitive grant program to enact preferred education reform policies, which included adoption of high-quality common standards (which could be demonstrated by participating in a consortium of states) and new assessments aligned to those standards (DOEd, 2009).

Common standards and common assessments, among other things, would address the variation in the stringency of state standards. And, common standards aligned with common assessments were expected to greatly enhance the interpretability of achievement results within and across states. The hope was that if states adopted the CCSS and signed on to one of the state assessment consortia (Smarter Balanced or PARCC), then policy makers, schools, and parents could finally gauge how their students were performing relative to their peers in others parts of the country. The same goals were behind the decisions of the consortia of states that developed common assessments for use with students with severe disabilities that would be assessed against alternative achievement standards derived from the Common Core standards (e.g., Dynamic Learning Maps and the National Centers State Collaborative).

Perfect comparability in testing, however, is not achievable (NRC, 1999a). From the inception of the Common Core assessments in 2010, questions arose about whether results within each consortium, let alone across consortia, were comparable when students were taking the tests via paper and pencil or computer, were using different electronic devices, were tested on different dates stretching over a multiweek administration window, or were subject to different accommodation policies (Hess, 2014). Moreover, well before the advent of Common Core assessments, NAEP was facing issues of comparability because states use different procedures for inclusion, accommodations, and so forth (NRC, 1999b). But such issues were not generally considered reason enough to abandon the idea of having common assessments that could provide comparable results across states.

The U.S. Department of Education incentivized states to adopt the Common Core standards and assessments but soon there was backlash: some felt that federal involvement in education had gone too far. In an effort to take back some local control over assessment, states started backing away from Common Core assessments. Relative to its predecessor version of the law, the Every Student Succeeds Act of 2015 allows states more flexibility in designing their accountability systems. Some of the possibilities include using the Common Core standards and assessments (Smarter Balanced, PARCC, or ACT Aspire), having their own customized state standards and tests, using one of the nationally recognized college entrance exams (ACT or SAT) as their high school assessment, or even giving districts within a state a menu of assessments to choose from.<sup>5</sup>

Over time, fewer states have been administering Common Core consortia assessments in their entirety as intended, and more states are moving toward creating their own unique assessment systems that include a blend of shared and customized elements (Marion, 2017). At their inception in 2010, 44 states and the District of Columbia joined either Smarter Balanced or PARCC; in spring 2019 only 15 states and the District of Columbia administered PARCC or Smarter Balanced and many of them did not do

---

<sup>5</sup> State Responsibilities for Assessments & Locally Selected, Nationally Recognized High School Academic Assessments. 34 C.F.R. § 200.2-3 (2016).



so for high school (Gewertz, 2019; Robelen, 2010). Some states are creating state tests using a combination of Common Core assessment items and their own state-customized items, some are partnering with vendors such as Pearson and Cambium Assessment (formerly American Institutes of Research Assessment) to develop their own tests, and some are using consortia tests in grades 3 through 8 and the ACT or the SAT at high school. In addition to choosing their assessment and vendor, states also define achievement levels differently.

*There is a trade-off, however, between variability and comparability.* At what point are comparisons between state test results no longer defensible? To what extent can states that modify assessment content and/or procedures continue to use the consortia's validity studies to support claims about the validity of their own state uses of the assessment? At what point in the modification of content and/or procedures does a state's use of the consortia's score scale become no longer meaningful?

As observed by Haertel and Linn in 1996, when examining issues of comparability in the context of performance assessment,

Different aspects of comparability will be more or less relevant in a given situation. As with any psychometric desiderata, the stringency of comparability requirements will depend on the kind of decision being made (e.g., "absolute" decisions about status with respect to a cutting score versus "relative" decisions about the rank ordering of students or schools); the importance of the consequences attached to those decisions; the level of aggregation at which scores will be reported and used (individuals versus aggregates like classrooms, schools, or states); the relative costs of mistakenly passing versus mistakenly failing an individual; the quality of other relevant, available information and how it is combined ... and the ease with which faulty decisions can be detected and revised. (p. 60)

The same principles apply to the current assessment context. This volume seeks to inform the design and use of large-scale assessments to help support intended inferences and actions. Chapter authors, who are all experts in educational assessment, examine the most pressing comparability issues in the current assessment system context and provide suggestions for moving forward. However, before turning to the comparability issues discussed in this volume, we first offer two critical definitions: (1) comparability and (2) assessment system.

### DEFINITION OF COMPARABILITY

Users of educational assessments assume that students' scores can be validly compared—they assume *score comparability*—even if those scores come from measurements taken at different times, in different places, or using variations in assessment content and procedures. Ideally, users could be assured that students with the same score possessed the same level of proficiency with respect to the domain of knowledge and skills a test was intended to measure (AERA, APA, & NCME, 2014).

Broadly speaking, there are at least three ways actual test scores necessarily fall short of this ideal. *First*, scores are imprecise—various sources of measurement error affect scores, introducing random error that limits score interpretations. *Second*, with few exceptions, the knowledge and skills a test actually measures do not perfectly

match the range of knowledge and skills that test users wish or intend to measure. *Third*, a range of influences can give rise to systematic differences in scores (i.e., differences in “expected scores”) among students who in fact possess equal proficiency with respect to the qualities the test actually measures. This third kind of imperfection, systematic influences that differentially affect scores of different examinees, comprises threats to score comparability and can arise from many sources. These three kinds of limitations can interact in complex ways, but, by and large, the first two—random errors and imperfections in the scope of knowledge and skills measured—on average affect all students’ scores in the same way. The third kind of limitation—factors affecting comparability—introduces systematic distortions that may affect score patterns across individuals or groups. This kind of limitation is the primary focus of the present volume. The following brief discussion is by no means exhaustive but is intended to clarify the scope of these comparability concerns addressed by the papers in this volume, and the importance of doing so.

Most obviously, if test performance requires proficiencies irrelevant to the knowledge and skills the test is intended to measure, and if some students’ performance suffers due to lack of those irrelevant proficiencies, then the scores of those students are not comparable to the scores of other students. This is a comparability concern because it systematically affects the scores of some students differently from others. On tests intended to measure knowledge and skills other than language proficiency per se (e.g., mathematical computational skills), scores of students hampered by limited language proficiency may be depressed for reasons unrelated to the construct the test is intended to measure. For assessments administered on digital platforms, if some students are unfamiliar with the technology employed, a similar issue may arise. Closely related to issues of irrelevant skill demands are issues of test bias. If item content is more interesting or more familiar to one or another identifiable group of students, score comparability may be compromised.

Threats to comparability may also arise due to differences in test administration or scoring conditions. The scores being compared may have been obtained using different test forms or may be based on different scorers’ judgments of students’ responses. Students may take digitally administered items on different kinds of devices or use test forms administered at substantially different times during the academic year. Score comparability may also be compromised if students in one jurisdiction perceive a test as “high stakes” and those in another jurisdiction do not, giving rise to differing levels of effort and engagement. In some cases, test administration conditions are deliberately altered to enable more valid measurements of target constructs for students requiring testing accommodations. Although appropriate accommodations can undoubtedly improve score comparability, sound and defensible use of testing accommodations can be challenging. Many of these threats to comparability are amplified when comparisons are made across different assessments and assessment systems.

When scores are compared for groups of students, comparability also demands that the groups be defined consistently, with proper attention to sampling, rules for exclusions and exemptions, and retesting practices.

Additionally, the issue of score comparability requires attention to the inferences drawn from test scores. Consider this scenario: A new, high-stakes test is introduced. Students are retested annually, and, over the first 2 or 3 years the test is in place, average



scores rise dramatically. If the intended inference as to the meaning of the test scores was limited to proficiency with respect to the content sampled on the test, demonstrated in just the ways the test called for, then one might validly infer that the rising scores showed proficiency increasing from year to year. If, however, the test scores are interpreted as indicators of a proficiency with respect to the broader domain of content the test was designed to represent, including both sampled and unsampled content, then the same pattern of rising scores might be attributed, at least in part, to realignment of curriculum and instruction to tested content elements at the expense of untested content elements. From the perspective of that broader intended inference, first-year and subsequent-year scores might not be entirely comparable. As these examples show, comparability is contingent on arguments and evidence about the intended purposes and uses of the test scores being compared. Comparability may be adequate for one interpretive purpose but not another.

### DEFINITION OF ASSESSMENT SYSTEM

Throughout this volume, the term *assessment system* is used and we need to be clear about the meaning and scope of this term as used in this volume, especially with respect to other discussions in the broader educational assessment literature (e.g., Herman, 2016; NRC, 2001, 2006). In general, an assessment system implies the existence of multiple assessments designed to function together to fulfill specific interpretive goals and purposes. The assessment system may be composed of assessments that range in form and content from teachers' classroom quizzes and midterm or final exams, to district, national, or international standardized tests. Whatever the specific tests included, the overarching purpose of the collective set of assessments making up the system should be to provide information that serves to promote student learning (e.g., Herman, 2016; Wiggins, 1998). The focus in this volume is on comparability concerns involving assessments that are primarily distal to the classroom—district, state, national, and international assessments.<sup>6</sup>

As noted by Coladarci (2002), "a collection of assessments does not entail a system any more than a pile of bricks constitutes a house." Rather, an assessment system is an assemblage adhering to principles that ensure that the elements are complementary and work together. In the National Research Council (NRC) report *Knowing What Students Know: The Science and Design of Educational Assessment* (NRC, 2001), three major system properties were described: *comprehensive*, *coherent*, and *continuous*.

*Comprehensive* means that a range of approaches is used to provide a variety of evidence to support educational decision making. Using multiple types of assessments and indicators that span the ways that a subject is expressed in the curriculum, and in typical instructional practices, enhances the validity and fairness of the inferences drawn by giving students various ways and opportunities to demonstrate their competence.

For the system to support learning, it must also have the property of *coherence*. One dimension of coherence is that there is consistency in the conceptualization of student learning underlying the various assessments within the system. While a state-level

---

<sup>6</sup> We are not suggesting that this restriction to the definition of assessment system should be broadly employed, just that we are focusing on more summative assessments in this volume.

assessment might be based on a model of learning that is broader and thus less fine grained than the model underlying the assessments used in classrooms, the conceptual base for a state assessment should be the same as that guiding assessment at the classroom level. In this way, results from assessments external to the classroom will be consistent with the more detailed understanding of learning underlying classroom instruction and assessment. The NRC (2006) also discusses the important property of vertical coherence, whereby the different levels of assessments conceptually align with curriculum and instruction at the given grade or academic level.

Finally, an ideal assessment system would be designed to be *continuous*. That is, assessments would measure student progress over time. To provide such pictures of progress, multiple sets of observations over time must be linked conceptually so that change can be observed and interpreted. Models of student progress in learning should underlie the assessment system, and individual assessments should be designed to provide information that maps back to the progression. Thus, continuity calls for alignment along the dimension of time.

Much of what concerns the chapters in this volume are assessments that have been designed for use at levels that are relatively distal in time and space from ongoing classroom instructional and assessment practice. The inferences made about student learning based on such distal assessments require levels and forms of comparability that are typically less critical for the highly contextualized interpretive uses associated with formative and summative classroom purposes.

Unless otherwise indicated, “assessment system” throughout this volume is therefore meant to apply to the types of systems designed to operate outside the classroom interpretive context.<sup>7</sup> It refers to a collection of assessments designed and used to measure student achievement with respect to some common content framework. In addition to the assessments themselves, an assessment system also refers to (1) the rules and policies governing uses of those assessments, (2) the infrastructure required to administer the assessments and to acquire and score students’ responses, and (3) the associated reporting structures and associated professional development designed to help users (i.e., students, teachers, parents, educational administrators, and policy makers) interpret the results. An assessment system may serve as the foundation for an accountability system that employs test scores, usually in conjunction with other kinds of information, to quantify the performance of students, schools, or districts and possibly to determine rewards or sanctions. As used here, however, “assessment system” is limited to the mechanisms for measuring and reporting student achievement to promote student learning and does not include the additional data sources and decision rules incorporated in an accountability system. However, at points we do make reference to the links and tensions between an assessment system and its accompanying accountability system.

This volume’s working definition of an assessment system is motivated by the high-stakes accountability context of K–12 education and testing. As used here, in addition to academic achievement tests, “assessment system” also encompasses tests of English language proficiency (including initial screening tests) used to classify students

---

<sup>7</sup> Such assessment systems may be within an individual school district or state, may span multiple states, or may span countries.

as English learners or as fully English proficient. It excludes, however, assessments of classroom climate and measures of socioemotional learning, important as these may be. Indicators of various student demographic variables may be used to report student achievement according to racial/ethnic group, gender, socioeconomic status, student language background, or other categories, but in this volume, these demographic variables are not treated as part of the assessment system itself. Also excluded are indicators of opportunity to learn (OTL), although consideration of OTL and related contextual factors may be essential if certain test score interpretations are to be fair and useful.

**Content framework.** The content framework undergirding an assessment system describes, in greater or lesser detail, what is to be taught and learned through formal schooling. At some level of abstraction, all of the assessments within a single assessment system can be linked back to a single, common framework, such as the CCSS. On closer examination, however, there may be multiple content definitions at various levels of specificity. The foundational document may set forth broad instructional goals but is unlikely to provide sufficient detail to guide either classroom instruction or the design of assessments. The CCSS, for example, is explicitly *not* a curriculum framework or test specification. Various intermediate documents may elaborate on the overarching framework. Some may prescribe the scope and sequence of instruction, and others may include “test blueprints” prescribing the mix of item types and content elements in particular assessments. The same assessment system may serve classrooms in which various textbooks are used for a given subject at a given grade level. These different textbooks may differ somewhat in the content and organization of instruction they prescribe, and, of course, individual teachers may adapt curriculum materials in different ways.

**Types of assessments.** The assessments within an assessment system may span multiple grade levels and subject areas. They may include specific assemblies of items used together (fixed-form tests), item assemblies created dynamically from calibrated item pools (computer adaptive testing), or both. Typically, there will be multiple forms of any given test for use over time (e.g., annual testing), as well as special forms for students requiring accommodations. Assessments may include multiple-choice items, other forms of selected-response items, constructed-response exercises, performance tasks, or various mixtures of these or other item formats.

**Comparability and context.** If an assessment system is to provide accurate, fair, and useful information to meet the needs of various audiences, it must be carefully designed to work within a given context. Alignment with content frameworks is fundamental to meeting virtually all such information needs. Users of information from an assessment system will appropriately assume that test scores reflect students’ mastery of significant content, going beyond the answers to specific questions actually administered. Alignment is essential if content frameworks are to provide trustworthy guidance as to the meaning of test scores.

In addition to alignment with content frameworks, many uses and interpretations will depend on the *comparability* of scores across students, across student groups, across

schools, across years, and, in some cases, across different kinds of assessments included within the assessment system. Clearly, not all assessments within an assessment system can or need to be directly comparable. There is not a requirement for a common scale for scores from all of the constituent assessments. It should also be noted that comparability is a matter of degree. At one extreme, scores from alternate forms of the same test might meet stringent psychometric requirements for *equating*, a fine tuning of score scales from different test forms that renders their scores entirely comparable. At the other extreme, there may be no common scale connecting a teacher's informal classroom assessment, used formatively to guide instruction, and the end-of-year, external summative assessment covering the same content, even though scores on those two very different tests would probably be positively correlated.

Forms and degrees of comparability for different purposes are complex and resist easy categorization. To give just a few examples, absent some compelling rationale, achievement standards defining (for example) "proficient" should be established in such a way that aggregate proportions designated as "proficient" do not change erratically from one grade level to the next, nor should they be grossly disproportionate across subject areas. If an assessment system offers the choice, for some assessment, of paper-and-pencil versus computer-based testing, or, more generally, a choice among digital platforms for computer-based assessments, then in order for the obtained scores to be reportable on a common scale, they should meet stringent standards for comparability. If scores for a certain demographic subgroup are to be compared across jurisdictions, those subgroups should be defined everywhere in the same way. To the degree possible, scores from students tested with accommodations should be reportable on the same scale, and interpretable in the same way, as for students tested without accommodations. These and other comparability issues are discussed throughout this volume.

## COMPARABILITY ISSUES ADDRESSED IN THIS VOLUME

As noted above, this volume is an attempt to provide guidance to key stakeholders, including state-level educational assessment and accountability decision makers, leaders, and coordinators; consortia members; technical advisors; vendors; and the educational measurement community regarding *how much* and *what types* of variation in assessment content and procedures can be allowed, while still maintaining comparability across jurisdictions and student populations. The volume also provides guidance and caveats to policy makers using assessment data to enact legislation, regulations, and district- and school-level guidance and also provides greater context for media using test scores to make comparability determinations. Here we briefly summarize the comparability issues addressed in this volume.

**Comparability of Individual Students' Scores on the "Same Test" (Chapter 2).** While comparability is often thought of as comparability across states or different tests, the first chapter in this volume begins by grounding the reader in comparability issues in the interpretation of a single test score of a single student. Charles DePascale and Brian Gong explain that while on large-scale assessments, individual student test scores on the same test are expected to be interchangeable (i.e., the student would be expected to receive the same test score if they took a different form of the test or took

the test under different conditions), meeting this goal is challenging. The term “same test” refers to various cases in which students may take different sets of items under different conditions. This chapter addresses how to evaluate whether comparability across forms and/or conditions is sufficient to support a particular inference or test use. Intended comparability may be supported through careful design decisions and psychometric procedures. There are also external threats that might affect the accuracy and/or interpretation of students’ scores. Students’ opportunity to learn the content assessed and familiarity with the item formats and tools used on the assessment are two types of comparability threats related primarily to their prior experiences. Threats to comparability that may arise from differences in the intended uses of the assessment and from different assessment contractors’ processing of the “same test” are also discussed. The process of establishing the comparability of individual student scores on the same test involves compiling sufficient evidence to support inferences and actions related to student performance based on those test scores.

**Comparability of Aggregated Group Scores on the “Same Test” (Chapter 3).** After examining individual students’ scores in Chapter 2, Leslie Keng and Scott Marion address the considerations and challenges associated with comparing scores from the same test at the aggregate level, such as between student groups, schools, districts, and states. While many principles and methodological approaches are similar to those addressed in Chapter 2, comparisons of aggregated group scores also must include, among other things, differences across jurisdictions in test delivery platforms, modes of administration, and testing accommodation policies. Since comparability is essential for establishing the validity of inferences, and validity is evaluated in the context of specific purposes and uses, this chapter explores the various uses and purposes associated with comparisons of aggregate performance for tests considered essentially the same; the categories of aggregate measures, or derived scores, used to compare group-level performance; and factors that can affect aggregate-score comparability. Because comparability exists on a continuum, the authors propose criteria that can be used to determine whether the preponderance of evidence supports comparability claims for an intended aggregate-score use or purpose and conclude with a practical framework for evaluating and mitigating threats to the comparability of group scores in current policy and practical contexts.

**Comparability Within a Single Assessment System (Chapter 4).** Mark Wilson and Richard Wolfe address comparability issues that arise within a single assessment system, focusing on summative results for individuals and aggregates (classrooms, schools, districts, and states). This chapter examines the validity of comparisons across grades, subjects, and years, and in interim results where they are strongly aligned to summative tests. The authors address the question of whether the different parts of the system measure the same or similar variables. As the authors note, test-to-test concordances only are useful or valid if there is confidence that the tests are addressing essentially the same underlying variables; as such, the chapter examines the alignment of subject-matter content, the design of the measurement constructs within the system, and the stringency of the different tests within the system. In essence, are the tests aligned and designed to attend to their intended uses? The chapter also addresses the reliability of the tests with respect to different uses and different levels of aggregation,



as well as the need for transparency in the system (i.e., what information should consumers have available to make decisions, and what level of technical documentation is needed to ensure that a system can be fully reviewed by expert evaluators).

**Comparability Across Different Assessment Systems (Chapter 5).** In this chapter, Marianne Perie expands the discussion beyond one assessment system and examines comparability issues when interpreting scores across more than one large-scale assessment. Policy makers want to compare performance across states and districts, using measures that go beyond NAEP. For instance, as policy has moved to focus on college readiness, there is also a desire not only to compare tests and state assessments across consortia but also to compare the results of such tests with traditional college admissions tests such as the ACT and the SAT. And, there is interest in international comparisons of state assessments to multinational tests such as the Trends in International Mathematics and Science Study (TIMSS) and the Programme for International Student Assessment (PISA). This chapter examines how different assessment systems might address score comparability of students, schools, districts, and states. Specifically, the focus is on elements of assessments required for comparability, understanding score comparability at different levels of aggregation, and psychometric constraints on desired inferences about students and schools across states and countries.

While issues pertaining to EL students, students with disabilities, and students from nondominant linguistic and cultural backgrounds permeate the volume, the steering committee determined that in addition to attention within chapters, comparability issues for these groups should also be the foci for individual chapters. As such, Chapters 6 through 8 address EL students, students with disabilities, and students from varying linguistic and cultural backgrounds as described below.

**Comparability When Assessing English Learner Students (Chapter 6).** In this chapter, Molly Faulkner-Bond and James Soland identify several decisions and test-score uses specific to EL students in the United States and introduce potential comparability issues concerning generalizations or comparisons about this population of students. These issues begin at the level of defining the population itself; because ELs are identified on the basis of test-based processes, decisions about who belongs in this subgroup, as well as reclassification criteria, may lack comparability across settings. Within the EL subgroup, comparing and interpreting English language proficiency scores is challenging due to differences in how tests are developed and scored, how states weight various subscores, and even how the construct of language proficiency is operationalized across measures. Achievement test score comparisons between ELs and non-ELs may be distorted by potential confounds between language and academic ability. Furthermore, many ELs take achievement tests using accommodations that complicate comparisons if not properly addressed, and ELs can also be part of other subgroups like students with disabilities that necessitate additional accommodations. Finally, using scales to estimate and compare growth for ELs (including comparisons to growth for non-ELs) is complicated by the shifting nature of the EL subgroup. In the chapter, the authors present several considerations for minimizing threats and supporting valid score use, both within and across populations and systems.

**Comparability When Assessing Individuals with Disabilities (Chapter 7).** Standardized testing procedures are meant to provide a level playing field for all examinees with respect to content tested, test administration procedures, and scoring processes. However, in some cases, aspects of standardized procedures may prevent examinees with disabilities from fully demonstrating their proficiencies. In such cases, accommodations may enable individuals with disabilities to better demonstrate what they know and can do. In this chapter, Stephen Sireci and Maura O’Riordan describe the various types of accommodations provided on statewide and college admissions tests, the resulting issues in score comparability, and how to evaluate the effects of test accommodations. The authors also examine test development procedures that may help make educational tests more accessible to individuals with disabilities, thereby reducing the need for accommodations.

**Comparability in Multilingual and Multicultural Assessment Contexts (Chapter 8).** Kadriye Ercikan and Han-Hui Por examine the impact of score comparability for students from different linguistic and cultural backgrounds on the validity of inferences from assessments. In addition to comparability issues arising in the context of international assessments given in multiple languages, the issue of consistent score meaning is also a concern for countries with populations from diverse language and sociocultural backgrounds, including countries with large immigrant populations. Recognition of the diversity within the United States led states to develop assessments in multiple languages and provide language tools and accommodations. This chapter highlights the complexity of comparability issues when tests are administered in multiple languages to students from diverse backgrounds and provides recommendations for optimizing comparability of adapted versions of tests.

**Interpreting Test-Score Comparisons (Chapter 9).** The concluding chapter of the volume, authored by Randy Bennett, is a cross-cutting chapter that examines—with all of the caveats and warnings described in prior chapters—how to best interpret test scores. And, as is likely evident by now, getting meaning from test results requires some type of comparison, be it to other test takers, oneself, or some absolute standard. Comparisons are strongest when the same measure is given under substantively the same conditions to comparable student samples at the same point in time. Comparisons become weaker as the measure, the assessment conditions, student samples, and the time of administration diverge. This chapter addresses when conditions are substantially the same as well as when divergence can occur. With respect to good practice, it is well to note that comparative claim statements can appear (or be implied) in score reports, press releases, websites, and other communications. When making such statements, it is best to determine first whether the same test is being used and if it is administered under the same conditions to comparable student samples at the same point in time. If not, the divergence(s) should be identified and a logical rationale for making the comparison should be articulated. The strength of the comparative claim should be adjusted as a function of (1) the extent to which the instruments, assessment conditions, student samples, and time between administrations diverge, and (2) the extent of the logical and empirical support available to back the claim and technical assistance committee review of this support. This chapter explores comparative claims

across this spectrum and suggests adjustments in terms of level of confidence based on either of these two factors.

As the chapters in this volume show, issues of comparability of assessment results are numerous and challenging but they are not insurmountable. It is our hope that, by surfacing these issues across a range of contexts where comparisons are inevitable, and often critical for informing policy and decision making, such comparisons can be approached in ways that are appropriate and useful. Each of the chapters offers cautions with respect to the types of comparisons of assessment results that are typically desired while also offering recommendations that can lead to more valid and useful inferences for those contexts of use that in turn can support equity, fairness, and enhancement of educational opportunities and outcomes.

## REFERENCES

- AERA (American Educational Research Association), APA (American Psychological Association), & NCME (National Council on Measurement in Education). (2014). *Standards for educational and psychological measurement*. Washington, DC: AERA.
- Coladarci, T. (2002). Is it a house...or a pile of bricks? Important features of a local assessment System. *Phi Delta Kappan* 83(10), 772–774.
- DOEd (U.S. Department of Education). (2009). *Race to the Top program executive summary*. Washington, DC: U.S. Department of Education. Retrieved from <https://www2.ed.gov/programs/racetothetop/executive-summary.pdf>.
- DOEd. (2010). *U.S. Secretary of Education Duncan announces winners of competition to improve student assessments*. Retrieved from <https://www.ed.gov/news/press-releases/us-secretary-education-duncan-announces-winners-competition-improve-student-asse>.
- DOEd. (n.d.). *Awards: Race to the Top assessment program*. Retrieved from <https://www2.ed.gov/programs/racetothetop-assessment/awards.html>.
- Gewertz, C. (2019, April 9). Which states are using PARCC or Smarter Balanced? *Education Week*. Retrieved from <http://www.edweek.org/ew/section/multimedia/states-using-parcc-or-smarter-balanced.html>.
- Haertel, E. H., & Linn, R. L. (1996). Comparability. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 59–78). Washington, DC: National Center for Education Statistics. Retrieved from <https://nces.ed.gov/pubs/96802.pdf>.
- Herman, J. (2016). *Comprehensive standards-based assessment systems supporting learning*. The Center on Standards & Assessment Implementation, WestEd. Retrieved from [https://www.csai-online.org/sites/default/files/resources/4666/CAS\\_SupportingLearning.pdf](https://www.csai-online.org/sites/default/files/resources/4666/CAS_SupportingLearning.pdf).
- Hess, R. (2014, April 30). SBAC responds to my queries about the Common Core tests. *Education Week*. Retrieved from [http://blogs.edweek.org/edweek/rick\\_hess\\_straight\\_up/2014/04/sbac\\_offers\\_answers\\_to\\_my\\_queries\\_about\\_the\\_common\\_core\\_tests.html](http://blogs.edweek.org/edweek/rick_hess_straight_up/2014/04/sbac_offers_answers_to_my_queries_about_the_common_core_tests.html).
- Jochim, A., & McGuinn, P. (2016). The politics of the Common Core assessments. *Education Next*, 16(4).
- Kaestle, C. F. (1983). *The pillars of the republic: Common schools and American society, 1780–1860*. New York: Hill and Wang.
- Kaestle, C. F. (2012). *The testing policy in the United States: A historical perspective*. The Gordon Commission on the Future of Assessment in Education. Retrieved from [https://www.ets.org/Media/Research/pdf/kaestle\\_testing\\_policy\\_us\\_historical\\_perspective.pdf](https://www.ets.org/Media/Research/pdf/kaestle_testing_policy_us_historical_perspective.pdf).
- Kalk, J. (2019, December 11). *New statewide test results show achievement gap throughout Cedar Rapids Community School District*. Retrieved from <https://www.kcrg.com/content/news/Cedar-Rapids-schools--566098691.html>.
- Kirkman, A. (2020, January 3). Majority of South Bend schools do not meet federal expectations, new report states. *South Bend Tribune*. Retrieved from [https://www.southbendtribune.com/news/education/majority-of-south-bend-schools-do-not-meet-federal-expectations/article\\_44cfff2-2da0-11ea-a537-a7e687e54948.html](https://www.southbendtribune.com/news/education/majority-of-south-bend-schools-do-not-meet-federal-expectations/article_44cfff2-2da0-11ea-a537-a7e687e54948.html).



- Marion, S. (2017, January 3). What's next for the Common Core and its assessments? *Future Ed*. Retrieved from <https://www.future-ed.org/whats-next-for-the-common-core-and-its-assessments>.
- Moss, P. A., Pullin, D. C., Gee, J. P., Haertel, E. H., & Young, L. J. (Eds.) (2008). *Assessment, equity, and opportunity to learn*. New York: Cambridge University Press.
- NAEd (National Academy of Education). (2009). *Education policy white paper on standards, assessments, and accountability* (L. Shepard, J. Hannaway, & E. Baker, Eds.). Washington, DC: Author.
- NRC (National Research Council). (1999a). *Uncommon measures: Equivalence and linkage among educational tests* (M. J. Feuer et al., Eds.). Washington, DC: National Academy Press. Retrieved from <https://www.nap.edu/read/6332/chapter/1>.
- NRC. (1999b). *Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress* (J. W. Pellegrino, L. R. Jones, & K. J. Mitchell, Eds.). Washington, DC: National Academy Press. Retrieved from <https://www.nap.edu/read/6296/chapter/1#ii>.
- NRC. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.
- NRC. (2006). *Systems for state science assessment* (M. Wilson & M. Bertenthal, Eds.). Washington, DC: The National Academies Press.
- OTA (Office of Technology Assessment). (1992, February). *Testing in American schools: Asking the right questions* (OTA-SET-519). Washington, DC: U.S. Government Printing Office.
- Resnick, D. P. (1982). History of educational testing. In *Ability testing: Uses, consequences, and controversies* (Vol. 2, pp. 173–194). Washington, DC: National Research Council.
- Rimel, A. (2019, September 21). Oregon dips in standardized test scores, mixed bag for mid-valley. *Corvallis Gazette-Times*. Retrieved from [https://www.gazettetimes.com/news/local/oregon-dips-in-standardized-tests-mixed-bag-for-mid-valley/article\\_cc900868-a925-59ce-923e-14af4ce01da3.html](https://www.gazettetimes.com/news/local/oregon-dips-in-standardized-tests-mixed-bag-for-mid-valley/article_cc900868-a925-59ce-923e-14af4ce01da3.html).
- Robelen, E. W. (2010, September 2). Two state groups win federal grants for Common tests. *Education Week*. Retrieved from <https://www.edweek.org/ew/articles/2010/09/02/03assess.h30.html>.
- Ryan, K. (2019, August 27). *Maryland's PARCC results show dip in math, improvements in English*. Retrieved from <https://wtop.com/maryland/2019/08/marylands-parcc-results-show-dip-in-math-improvements-in-english>.
- Sethrie, J. (2020, January 6). Minnesota report card: Small schools score higher. *Fillmore County Journal*. Retrieved from <https://fillmorecountyjournal.com/minnesota-report-card-small-schools-score-higher>.
- Symposium: Equity in educational assessment. (1994). *Harvard Educational Review*, 64(1).
- Tyrrell, J. (2019, April 4). *Survey: 45% of test-takers boycott ELA exam*. Retrieved from <https://www.newsday.com/long-island/education/schools-ela-opt-outs-test-boycott-1.29381145>.
- Vinovskis, M. A. (2019). What use is educational assessment? In A. I. Berman, M. J. Feuer, & J. W. Pellegrino (Eds.), *The Annals of the American Academy of Political and Social Science*, 683, 22–37.
- Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. San Francisco, CA: Jossey-Bass.

# Comparability of Individual Students' Scores on the "Same Test"

Charles DePascale and Brian Gong,  
*National Center for the Improvement of Educational Assessment*

## CONTENTS

INTRODUCTION .....	25
Interpreting an Individual Test Score: Comparability and Validity .....	26
Purposes for Comparing Test Scores .....	26
The Same Test .....	27
APPROACHES TO PRODUCING COMPARABLE TEST SCORES .....	29
Design Approaches to Producing Comparable Test Scores .....	30
Psychometric Approaches to Producing Comparable Test Scores .....	33
THREATS TO COMPARABILITY .....	38
Opportunity to Learn .....	38
Familiarity with Item Formats and Tools Used on the Assessment .....	41
Differences in Intended Uses of Test Results .....	41
Differences Among Assessment Contractors .....	43
CONCLUSION .....	45
REFERENCES .....	47

## INTRODUCTION

In large-scale assessments, individual student test scores on the same test are expected to be comparable, but meeting this goal is challenging. The challenge is exacerbated in large-scale K–12 testing because the term “same test” refers to various cases in which students may take different sets of items under different conditions. This chapter addresses how to evaluate whether comparability across conditions is sufficient to support a particular inference or test use. Common threats to comparability arise from a lack of attention to design decisions and psychometric procedures. There are also external threats that might affect the accuracy and/or interpretation of students’ scores. Students’ opportunity to learn (OTL) the content assessed and familiarity with the item formats and tools used on the assessment are two types of comparability threats related primarily to their prior experiences. The process of establishing the comparability of individual students’ scores

on the same test involves compiling sufficient evidence to support inferences and actions related to student performance based on those test scores.

### **Interpreting an Individual Test Score: Comparability and Validity**

Questions about the comparability of individual students' test scores begin not with the comparison of the scores of two or more students, but with the interpretation of a single test score of a single student.

With regard to an individual student's test score, a fundamental comparability question is whether the student would receive the same test score if they took a different form of the test or took the test under different conditions, that is, whether the test scores are interchangeable. If the test is computer based, would the student have received the same score with a different delivery platform, with a different response device, or on a paper-and-pencil form of the test? Would the student have received the same score with a different set of accommodations?

At its core, the question of comparability is a question of validity (Winter, 2010). What inferences about a student's knowledge and skills can we draw from the test score? What claims about a student's performance are supported by the score? What are the appropriate interpretations and uses of the score? These are the underlying questions that must be answered when we attempt to determine the comparability of individual students' scores.

The inferences that we draw about an individual student's performance may be different if we limit our consideration of comparability to only those cases in which all students take a test under the same testing conditions. We can consider the same test form administered under two different conditions. In scenario A, all students are required to take the test under strict timing with a limited set of accommodations; in scenario B, all students are allowed to complete the test and a wide selection of accommodations are available. Each scenario is likely to result in different test scores for many students.

It is likely true that test scores from scenario A are not comparable to test scores from scenario B. It is not true, however, that one scenario produces *comparable* test scores and the other does not. In both scenarios, whether an individual student's scores are comparable depends on the questions being asked about student performance and the interpretations and inferences made about student performance based on the test score.

### **Purposes for Comparing Test Scores**

When we move beyond the interpretation of an individual student's test score to make an external comparison, two types of comparisons are common. One involves comparing an individual score to a fixed standard or a fixed point on a scale. The other involves the direct comparison of two individual students' scores.

Over the past 20 years, perhaps the most common comparison made with individual students' test scores is the comparison of a student's score to the threshold scores (i.e., cut scores and passing scores) that delineate performance standards on state assessments. On the basis of such comparisons, student performance is classified into an achievement level. Achievement level classifications are used for a variety of purposes, including providing information to parents and students about student achievement of

state standards, school accountability, educator accountability, and student promotion or graduation.

Classification of student performance to an achievement level is not the only reason for comparing an individual student's test score to a point on a scale. In addition to criterion-based performance standards, a student's test score might be compared to a norm-referenced standard such as the 20th, 50th, or 95th percentile to inform placement decisions or eligibility for particular programs. A student's test score might also be compared to an expected score signifying an acceptable level of growth or progress from one test administration to the next to determine whether the student is on track to reach a specified goal.

The direct comparison of two or more test scores for the same student or different students may seem quite familiar; however, such comparisons of individual students' test scores are far less common than comparisons to a criterion- or norm-referenced standard. One example of such a direct comparison involves rank ordering of test scores to identify the highest or lowest performing students on a particular test. Another example is when a teacher regards a particular student's performance as a norm or criterion against which other students' performance is compared.

In general, however, the direct comparison of two or more test scores is much more common with aggregate school-, district-, or state-level scores, which are discussed in Chapter 3, Comparability of Aggregated Group Scores on the "Same Test."

### **The Same Test**

Any discussion of comparability issues associated with scores on the same test has to begin with a common understanding of what is meant by the term "same test" in the context of large-scale assessment. We are well beyond the point when the measurement community and most policy makers would consider any two tests with the same title (e.g., "Algebra I" or "Grade 3 Reading") to be the same test. Additionally, agreement is widespread that building two tests to the same set of content standards is a necessary but not sufficient criterion for considering two tests to be the same test. For example, few would consider the "Grade 3 Reading Tests" implemented in 2015 by Smarter Balanced, the Partnership for Assessment of Readiness for College and Careers (PARCC), and other state assessment programs in response to states' adoption of the Common Core State Standards to be the same test. One must go deeper than the same set of standards and consider factors such as the test blueprint, item specifications, and administration conditions.

It is rare, however, to find the one example that most people would agree fits the definition of the same test: the case where the same set of items is administered to all students at the same time under the same testing conditions. Even traditional, fixed-form, common assessments that were the norm for state assessments under the No Child Left Behind Act of 2001 (NCLB), such as the Grade 3 Reading Test, often included embedded field-test items that were matrix sampled and varied across students. Although the items from which student scores were produced were the same across students, it would be inaccurate to declare that the students' entire testing experiences were the same. The transition to computer-based testing introduced a plethora of supports and tools available to all students that may further alter the test experience across students (PARCC, 2017; Smarter Balanced, 2018). With the advent of computer-based adaptive

testing, it may be the case that no two students in a state who are taking the Grade 3 Reading Test complete the same set of items.

What then do we mean when we refer to student scores on the same test? Using a state's Spring 2019 Grade 8 Reading Test as a reference, Box 2-1 describes what we consider to be the same test for the purposes of the discussion in this chapter.

Within the category labeled "same test" there is considerable variation in the set of test items that are completed by individual students. A defining characteristic of the test forms in this category, however, is that they are all constructed to meet the same test blueprint and test specifications. In practice, a state or testing company reports scores from each of these test forms on a single reporting scale and regards the individual student scores as interchangeable.

Test forms listed within the "gray area" have some similarities to the test forms that are considered to be the same test, but also have some key differences. One key similarity is that it is likely that results from these test forms are reported on the same scale as the original test and treated as interchangeable scores.

A "short form" of a test may be built to the same blueprint and even include the same item types, but it is likely less reliable than the original long form of the assessment. In many cases, however, the short form excludes certain item types or alters the distribution of items across item types. As we discuss in the section on psychometric

### **BOX 2-1** **Examples of Tests Considered to Be the Same as a** **State's Spring 2019 Grade 8 Reading Test**

#### **Same Test**

- All standard operational forms of the Spring 2019 Grade 8 Reading Test
  - Fixed form with embedded field-test items
  - Multiple fixed operational forms
  - Matrix-sampled operational forms
  - Computer adaptive test
- Standard forms administered with accommodations and nonaccommodation tools and supports
- Standard forms administered in different formats (e.g., paper based, computer based)
- Standard forms with items administered in random order
- Alternate forms of the standard Spring 2019 Grade 8 Reading Test to be administered at different times of the year (e.g., summer retest, fall administration if there was block scheduling)

#### **Gray Area**

- "Short form" of the Spring 2019 Grade 8 Reading Test
- "Focused retest" of the Spring 2019 Grade 8 Reading Test designed to determine only whether a student meets the mastery or proficient cut score
- Spring 2018 Grade 8 Reading Test
- Spring 2020 Grade 8 Reading Test

#### **Different Test**

- Spring 2019 Grade 7 Reading Test
- Grade 8 Reading Interim/Benchmark Assessment
- Spring 2019 Grade 8 Reading Test, released and administered by local choice after 2019

approaches to producing comparable scores, differences in reliability can affect score comparability and our inferences about individual student performance.

A "focused retest" such as the type administered in Massachusetts in the early 2000s to high school students attempting to meet a graduation requirement is designed to measure student performance at a more precise point on the reporting scale (i.e., at a significant achievement level cut point) and produce a comparable decision about student achievement at that point on the reporting scale. Even if the same test blueprint is applied to construct the test, it is likely that the original test and focused retest will differ in difficulty and reliability at various points along the reporting scale.

Versions of the Grade 8 Reading Test from the previous year or following year are included in the gray area category for a different reason. In many cases, those test forms can be considered alternate versions of the operational form administered within the same year. In other cases, however, key changes in the testing program from one year to the next have the potential to affect score comparability. Among these are changes in use of the test or stakes associated with individual student scores, transition to a new assessment contractor, and changes in achievement standards.

The three examples of test forms within the "different test" category have critical differences from the original test form that make it impossible to label them as the same test for our purposes. As was the case with the test forms in the gray area, it is possible that results from test forms in this category will be reported on the same scale as the original test form (including the use of a vertical scale). A test designed to measure performance at a different grade level (e.g., the Grade 7 Reading Test) though will have been aligned to different content standards and built to a different test blueprint. A released test form or an interim or benchmark version of the Grade 8 Reading Test may or may not be built to the same blueprint and test specifications as the original test form, but differences in the timing and conditions of the test administration will result in the same test score leading to different inferences about student performance. For example, it may be the case that an interim assessment built to the same blueprint as the original test form supports the same inferences about student performance at the time the interim test was administered (i.e., the student has met the proficient cut score), but it may provide different inferences about how the student is performing or will perform at the end of the school year.

## **APPROACHES TO PRODUCING COMPARABLE TEST SCORES**

Achieving comparability of individual students' scores on the same test requires very thoughtful and careful planning, execution, monitoring, and evaluation. Comparability is the result of a process that involves a combination of design and psychometrics (i.e., statistics). Comparable scores cannot be achieved solely through the application of formal psychometric procedures, nor can comparable scores be achieved solely through the design of a large-scale assessment system. Most importantly, the psychometric and design approaches to achieving comparability are interdependent. The psychometric procedures, particularly in the case of state assessment programs, depend on strict design assumptions having been met. The design approaches rely on psychometric procedures to account precisely for differences in difficulty between forms of the test.

The psychometric approaches to achieve comparability include a range of procedures designed to enable direct comparisons of student performance on different sets



of items, regardless of whether those sets of items are intended to be the same test or different tests. In the case considered here, in which different sets of items are intended to be the same test, the level of desired comparability is that scores from any of the alternate forms of the same test can be treated as interchangeable. In other applications, it may be sufficient to establish that classifications of student proficiency on the two assessments are comparable (e.g., determining a common college-readiness cut score on two different assessments).

The design approaches to achieve comparability involve a series of decisions related to the overall design of the assessment system. They include decisions related to how the test is developed, administered, and scored. They also include decisions about who will be taking the test and how the results will be used. As with the psychometric procedures, the goal with large-scale assessment programs that include multiple forms of the same test is to make design decisions that result in scores that are interchangeable across test forms.

### **Design Approaches to Producing Comparable Test Scores**

Design approaches to producing comparable individual student scores across alternate forms of the same test begin with an understanding of the construct being assessed. An understanding of the construct and the type of evidence needed to support the claims and inferences made about student performance on the construct are at the core of principled approaches to assessment design.

On large-scale K–12 assessments, the starting point for understanding the construct is usually the set of college- and career-readiness content and performance standards adopted by the state. These standards define the knowledge and skills that students are expected to have achieved at the end of a grade level, grade span, or course. Efforts to ensure alignment often begin with the development of evidence models or statements that describe in detail the aspects of student responses that would provide evidence needed to support the claims being made about student achievement based on performance on the test (Zieky, 2014). These evidence statements are supported by the development of detailed blueprints and test specifications that define what will be included on the test and how it will be measured. Test blueprints may contain information about the total number of items and points on the test, how those will be distributed across test sessions, and, most importantly, how the content standards to be assessed will be distributed across those items and points. Test specifications include additional information about the design of the assessment: details regarding item types, cognitive complexity (e.g., depth of knowledge), mode of administration (i.e., computer based or paper based), timing, and the nature and use of accommodations. In K–12 state assessments, the finished product is subjected to formal alignment studies that evaluate the degree to which each individual test item measures the standard or standards it was designed to measure and also the degree to which the set of items (i.e., the test form) measures the complete set of standards it claims to measure.

From a comparability perspective, this level of understanding of the relationships among the construct, the standards, and the assessment is critical to understanding what deviations from the “same test” are likely to impact the measurement of the construct and, consequently, impact the comparability of individual student scores.

These deviations include changes to the blueprint across test forms, such as shortening the test or changing the balance of selected-response and constructed-response items. Deviations that may impact the comparability of individual test scores also apply to the level of flexibility allowed within the administration of a single test form or alternate test forms built to the same test blueprint and specifications. Changes to the test blueprint such as shortening the test are addressed in the next section on psychometric approaches to producing comparable test scores. In the remainder of this section our focus is on decisions related to standardization versus flexibility that have the potential to impact the comparability of individual student scores on the same test.

### *Standardization Versus Flexibility*

Standardization has been a cornerstone of large-scale assessment. Content, administration, and scoring are the three pillars of standardization that drove the construction of large-scale assessment throughout the 20th century and into the beginning of the 21st century. One cannot overstate the importance attached to students taking test forms that contain the "same" content, are administered to all test takers under the same testing conditions, and are scored using the same specified scoring procedures. Standardization was considered essential to ensuring accurate measurement of student performance by controlling error and minimizing the impact of factors that are irrelevant to the construct being assessed or the purpose of testing and that might distort inferences about student performance. Standardization was also considered an essential requirement for making direct comparisons of individual students' scores.

With the Improving America's Schools Act of 1994 and NCLB, reauthorizations of the Elementary and Secondary School Act of 1965, federal assessment requirements affected the conception of standardization in several ways. Specifically, the requirements for (1) standards-based assessments aligned to challenging academic content standards, and (2) the inclusion of all students in large-scale assessments, resulted in a need to rethink what was meant by standardization with regard to key aspects of test administration and scoring. Within the category of test administration, two areas where more flexibility was allowed were in the timing of the test and in the use of test accommodations. Within the category of scoring, the increased use of constructed-response items and other item types that could not be easily scored by machine on large-scale assessments created a need for new sets of scoring protocols and procedures to ensure standardization in scoring, that is, to increase the likelihood that a student's response would receive the same score regardless of who scored it and when it was scored.

In each of these cases, the comparability of individual students' scores rests on the argument that the flexibility introduced into the testing process actually was more suited than strict standardization to minimizing error, removing irrelevant factors that might affect test performance, and ultimately producing more accurate student test scores.

**Timing** Historically, large-scale standardized tests were administered with time limits and strictly timed sections. Whether the test was regarded as a speed test or a power test, strict time limits were regarded as necessary to support the claim that all students



took the test under the same conditions, thereby meeting a necessary condition to support direct comparison of scores.

With the shift to standards-based assessment there was a prevailing belief that speed of response was not a component of the constructs being measured and that students should be provided adequate time to complete all items on the test in order to provide a more accurate estimate of their level of proficiency. In some states, the result was a timing policy in which timed test periods were expanded to levels designed to ensure that all students were able to complete the test. Test sessions were scheduled for 50 or 100 percent longer than the time expected for the vast majority of students to complete the test. In Massachusetts, the tests of the Massachusetts Comprehensive Assessment System are untimed, with the limitation that a test session must be completed within a single school day (MA DESE, 2019).

**Use of accommodations** Prior to the requirements of the Improving America's Schools Act, the Individuals with Disabilities Education Act, and NCLB to include all students in testing, it was not uncommon for 15 to 20 percent of students to be excluded from large-scale state assessments, a group including most students with disabilities (Lehr & Thurlow, 2003). The requirement to include all students in state assessments brought with it the requirement to allow students with disabilities the use of appropriate test modifications (i.e., accommodations) during testing. The 2014 Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014), referred to hereafter as the Joint Standards, define these test accommodations as adjustments that do not alter the construct being assessed that are applied to test presentation, environment, content, format (including response format), or administration conditions for particular test takers; and specify that accommodated scores should be sufficiently comparable to unaccommodated scores that they can be aggregated together. In practice, states often group accommodations into four categories:

1. Presentation, where accommodations involve varying the manner in which the test is presented (e.g., large-print, Braille, translated, read aloud);
2. Response, where accommodations involve varying the manner in which the student responds (e.g., use of graphic organizers, nonuse of a separate answer sheet, scribe);
3. Setting, where accommodations involve varying the setting in which the test is administered (e.g., one-on-one, small group); and
4. Timing, where accommodations involve varying the timing of the test (e.g., extended time, additional scheduled breaks).

The argument for increased flexibility and the use of accommodations for students with disabilities was centered on accessibility to the test and leveling the playing field for students with disabilities. Again, the goal was to produce a more accurate measure of the construct being assessed. One method recommended for determining if the accommodation leveled the playing field without affecting the construct required the demonstration of a differential boost. A differential boost indicates that the use of the accommodation affected the performance of students with disabilities for whom

the accommodation was intended more than it increased the performance of students without the disability.

In most cases, there was little debate over the appropriateness of several widely used and established accommodations such as Braille or large-print test booklets, small-group test administration, or allowing extended time. Other accommodations, such as use of a read-aloud accommodation on a reading or English language arts test or the use of a calculator in the noncalculator section of a mathematics test were often the subject of intense debate within and across states. In the case of the read-aloud accommodation, the debate often focused on whether the construct being assessed and the inferences drawn about student performance were related to decoding (i.e., being able to read the text) or comprehension (i.e., being able to interpret and apply the contents of a text). For a fuller discussion about the uses of accommodations, see Chapter 7, Comparability When Assessing Individuals with Disabilities.

**Scoring procedures** The increased use of constructed-response items and other item formats that could not be easily machine scored introduced variability into the scoring process that did not exist when tests consisted almost exclusively of machine-scorable, multiple-choice items. Not only were the rules for scoring a student response more complex, but also training and monitoring were required to ensure that groups of human scorers were able to apply those scoring rules consistently and accurately. The infrastructure needed to support the development of scoring rubrics, training materials, training, scoring, and real-time monitoring of scoring had to be developed and implemented to support the argument for comparability of individual student scores.

Scoring the same items across multiple test administrations within and across years added additional strain to the process of ensuring standardization in scoring. In addition to ensuring that there had been no changes to the item and scoring rules and that training and training procedures were consistent across years, it became necessary to take additional steps to ensure that student responses were being scored consistently from one administration to the next.

### Psychometric Approaches to Producing Comparable Test Scores

In large-scale assessment programs offering alternate forms of the same test, the desired level of comparability is that the scores on the alternate forms of the test are interchangeable. That is the case within state assessment programs when students are administered multiple forms of the Grade 8 Reading Test within the same year or when students take a new form of the Grade 8 Reading Test each year. That is also the case when students take multiple administrations of the ACT or the SAT as part of their college admissions process. In each of these cases, the task is to link the alternate forms of the test in such a way that an individual student would receive the same score regardless of which form is administered.

The type of linking required to achieve scores that are interchangeable is called *equating*. Although it is quite common to hear all forms of linking tests referred to colloquially as equating, the term is reserved for linkages that meet a strict set of requirements and where the intended interpretation is that the scores from the two tests can

be considered interchangeable for all intended purposes and uses. Holland and Dorans (2006) contrast equating with linking done for two other purposes: *predicting* and *scale aligning*. They refer to predicting as the oldest form of linking and one that is often confused with other methods of score linking.

As the name implies, the goal of linking approaches that are classified as predicting is to predict a student's score on a particular test from other information that is available about the student. That information might be a score on another test, which is where confusion with other purposes might arise. The other information about the student that is used as a predictor, however, could just as easily be performance on multiple tests, grades, dispositions, background characteristics, or anything that would improve the prediction of the student's score on the test of interest. Although a better prediction might be achieved by a predictor test that has a high degree of similarity to the test on which the score is being predicted, there is no requirement that the tests are similar in any way or even measure the same general content area, let alone the same construct. Regression procedures are the common tool to accomplish linkages for predicting.

The purpose of scale aligning, also referred to simply as scaling, is to produce comparable scores on two tests by transforming the scores from two different tests onto a common scale (Holland & Dorans, 2006). Holland and Dorans describe six types of scaling, categorized by whether the two tests are designed to measure similar constructs and, if so, whether the two tests have similar reliability. In each of the six cases, the result of scale aligning is that the scores from the two tests are placed on a common scale to produce comparable scores. In none of the cases classified as scale aligning, however, are the scores on the two tests to be considered interchangeable. That interpretation is reserved for linking procedures that meet the strict requirements necessary to be classified as equating.

In this chapter, we do not address all of the examples of scale aligning, but there are two that fall under the heading of calibration that are particularly relevant to the topic of comparability, large-scale state assessment, and tests that some might consider the same test. Calibration is considered a strong form of linking two tests and applies to situations in which the two tests meet many of the same requirements as equating, such as two tests that are designed to measure the same construct and may even be designed according to the same general test specifications (Dorans, Moses, & Eignor, 2010; Linn, 1993; Mislevy, 1992). The first is the case in which calibration is used to link a shorter form of a test to the longer, original form of the same test. The second is the general case in which item response theory (IRT) procedures have been used to place test items on a common scale.

With increasing concerns about the time needed to complete tests measuring complex, college-readiness standards, states are facing demands to reduce testing time by shortening tests. In these cases, most states attempt to maintain the same test blueprint in terms of content, cognitive demand, and the types of tasks that students are required to perform; however, even when such conditions are met, it is likely that the shortened test will have a lower level of reliability than the original test. Calibration procedures can be conducted to link the shorter and longer tests in a way that provides scores that have sufficient comparability to allow comparisons of scores of individual students on the short and long forms. It is also possible to apply calibration procedures to link the two tests so that comparisons can be made of aggregate group performances such as estimates of the percentage of students scoring above the proficient benchmark. It

cannot be assumed, however, that one set of calibration procedures will provide the same level of accuracy for individual and group comparisons (Linn, 1993; Mislevy, 1992). When it is desirable or necessary to treat long and short forms of a test as the same test, it is important to ensure that the appropriate linking procedures are used to support the most important comparisons and to understand how well all intended or likely comparisons can be supported.

A second caution is to not confuse equating with the use of IRT procedures to place items on a common scale. With the increased use of computer-based adaptive testing and states making use of items from commercial or shared item banks, there is a great reliance on the use of IRT to place items from many different tests and testing situations onto a common scale. It is true that applying IRT to place items on a common scale is typically a first step in equating large-scale assessments, but simply building alternate test forms by selecting items from a pool of items on the same scale should not be regarded as equating two tests. Even if all of the assumptions regarding the use of IRT to place the items on a common scale have been met, to use that common scale and those item parameters to claim that two tests built with items from that scale have been equated, all of the design and psychometric approaches to producing comparable test scores and threats to comparability discussed in this chapter must be considered.

### *Equating*

The purpose of equating tests is to allow the scores from each test to be used interchangeably, as if they had come from the same test. (Holland & Dorans, 2006)

We have established that in the vast majority of cases in large-scale assessments in which there is interest in the comparability of individual students' scores on the same test it is unlikely that those students took the same test in the literal sense (i.e., completed the same set of test items). Administering test forms with different items is the desirable case when referring to alternate forms of the same test administered in different years. Administering test forms which were not identical, however, was also necessary within the same year with fixed-form large-scale state assessments that included embedded field test items or matrix-sampled equating items, which were the norm for nearly two decades. In recent years, administering different sets of items across students has become increasingly common with the emergence of various forms of computer adaptive testing and the renewed interest in and use of matrix sampling to assess complex standards such as the Next Generation Science Standards.

When the goal is to treat individual students' test scores from two test forms that contain some different items as interchangeable then it is necessary to make a direct link between the two tests through equating. To support the claim that the test scores are interchangeable, equating has the strongest set of assumptions (i.e., requirements) of all of the approaches to linking two tests. Holland and Dorans (2006) identified five requirements for two tests to be equated successfully (see Table 2-1).

Each of the five requirements reflects, to some degree, a theoretical concept or measurement ideal that cannot be fully met in practice with real tests administered under real testing conditions, and with real people taking those tests. That, however, does not excuse test developers and test users from the need to adhere to best practices with regard to the development of test forms and the interpretation and use of individual

**TABLE 2-1** Requirements for a Linking to Be Considered Equating

The equal construct requirement	The tests should measure the same constructs.
The equal reliability requirement	The tests should have the same reliability.
The symmetry requirement	The equating function for equating the scores of Y to those of X should be the inverse of the equating function for equating the scores of X to those of Y.
The equity requirement	It should be a matter of indifference to an examinee to be tested by either of two tests that have been equated.
The population invariance requirement	The choice of (sub)population used to estimate the equating function between the scores of tests X and Y should not matter; that is, the equating function used to link the scores of X and Y should be <i>population invariant</i> .

SOURCE: Holland and Dorans (2006).

students' test scores. Striving to meet the five requirements, demonstrating an attempt to meet the requirements, and providing evidence of the extent to which the requirements have been met are critical to supporting an argument that individual students' test scores are not only comparable but are, in fact, interchangeable across test forms.

The starting point for the comparability of individual students' scores on the same test are the equal construct and equal reliability requirements. Stated simply, the two tests should be built to the same test specifications. Those test specifications must include factors such as the balance of representation of items (or score points) across content and cognitive processes, the use and distribution of items (or score points) across item types, and the number of items (or score points) that will be included on the test. In short, the test specifications should include details on any factors that could affect whether the two tests measure the same constructs and have the same reliability. Recall that in the section "Design Approaches to Producing Comparable Test Scores" we discussed several design choices where decisions must be made regarding whether a particular factor affects the constructs or claims being made about student performance (e.g., testing time, inclusion of particular item types, and use of accommodations). For the sake of this discussion, it is assumed that those issues are resolved before psychometric procedures are applied to equate the two test forms.

The equity requirement states that it should be a matter of indifference to an examinee to be tested by either of two tests that have been equated. In practice, we know that an individual student's test score might vary based on the particular set of items that they encounter on their test form. For example, if all items except the final item were the same across two test forms, it would make a difference to a student if they were able to respond correctly to the final item on form A but not to the final item on form B. The impact of an individual student being more or less familiar with a particular item or items on a test form will be lessened if the items are sampled from the domain in the same way across test forms and as the number of items and total points increase on each test form.

It is a greater concern to equity, however, if the likelihood of a student with a particular level of achievement having their performance classified as proficient or meeting a particular cut score varies across test forms. The purpose of equating is to adjust for

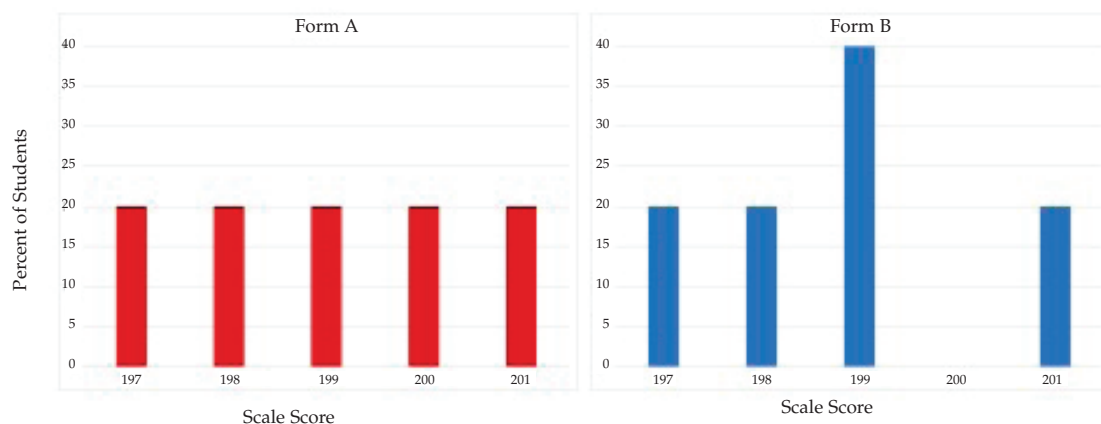


slight differences in difficulty across test forms so that an individual student will not be advantaged by receiving a test form that is slightly less difficult or penalized by receiving a test form that is slightly more difficult. However, equating, *per se*, cannot account for gaps in the reporting scale that might make it more or less difficult for students to attain a particular cut score based on the test form they receive.

Consider an example in which the proficient cut score on a state assessment is 200. Due to a combination of the difficulties of the particular items on form B and rounding rules adopted by the state it is possible for students to earn a scale score of 200 on form A, but it is not possible for students to earn a scale score of 200 on form B (see Figure 2-1). There is a gap in the reporting scale on form B such that students may earn a scale score of 199 or 201. In this example, the 20 students who would have earned a scale score of 200 on form A and a performance classification of proficient now receive a scale score of 199 and their performance is not classified as proficient. In this case, although the test forms are equated, it is clearly not a matter of indifference to students performing near the cut score of 200 which test form they receive.

Note that the example described here is not related to measurement error in the individual student's score. It will always be the case that there are students whose true score is near an achievement level cut who will score above the cut on one test administration and below the cut on a second administration—even if they were to take the same test form both times. This is more a case in which a student whose true score is near the achievement level cut is unable to earn a particular scale score because of the properties of the test form or the reporting scale.

The symmetry and population invariance requirements are relatively easy to evaluate in practice and the likelihood of meeting them is increased by meeting the equal construct and reliability requirements, that is, by developing test forms that measure the same constructs with the same level of reliability (Dorans et al., 2010). It is not likely that the equating functions produced to test either the symmetry or the population invariance requirements will be exactly the same based on individual test forms and samples of students, but it is important to understand where deviations from the requirements are occurring, to understand whether they fall within expected ranges



**FIGURE 2-1** Example of the difference in scale scores for two different forms of the same test.



given factors such as sample size and level or distribution of performance, and to attempt to determine why those deviations are occurring.

Assuming that the construct of the test forms is the same for all subgroups or samples of the population, demonstrating that the reliability of the test is consistent across subgroups is an important part of demonstrating that the test scores across forms are comparable for individual students.

There are a number of statistical approaches to conducting equating procedures and it is beyond the scope of this chapter to describe them or to discuss their advantages and disadvantages. With regard to the comparability of test scores, however, it is important to remember that all equating procedures begin with the requirements that the test forms being equated are measuring the same constructs with the same level of reliability. The statistical procedures will produce results even if those two requirements are not met, but those statistical results will not produce comparable test scores for individual students.

### THREATS TO COMPARABILITY

In describing design and psychometric approaches to producing comparable test scores for individual students, certain threats to score comparability are clearly implied. Test forms that are constructed to different test blueprints, include different item types, or are of significantly different length (i.e., differ in reliability) are unlikely to produce scores with a level of comparability needed to be considered interchangeable. Test forms that have significant differences in testing time, accommodations offered, and mode of administration (paper based versus computer based) are likely to have issues that must be examined and resolved before declaring that the scores they produce are comparable for individual students.

In addition to those internal threats to comparability (i.e., threats built into the assessment), there are additional threats to comparability that should also be understood when interpreting and using the results from large-scale assessments.

### Opportunity to Learn

OTL has long been acknowledged as a major instructional factor affecting student achievement (Kurz, 2011). At a gross level, OTL is defined in terms of the percentage of time in the school schedule allocated for instruction (Carroll, 1989). With regard to assessment, historically “OTL has been conceptualized as opportunity to learn what is tested” (Haertel, Moss, Pullin, & Gee, 2008).

More refined definitions include consideration of resources and other factors that affect the quality of instruction, students’ preparedness to participate in learning, and interactions among teachers, students, and standards (AERA et al., 2014; Banicky, 2000; Cooper & Liou, 2007; Friedlaender & Darling-Hammond, 2007; Shepard, Hannaway, & Baker, 2009). Among these factors are unequal or inequitable access to curriculum, instruction, and resources (including teacher and administrator quality); teaching to the test and other practices associated with high-stakes testing; adequacy of school finance; disciplinary and exclusionary practices; access to culturally responsive teaching and curriculum and school culture; access to evaluation for special needs; and implementation of 504 plans and individualized education plans.

With regard to large-scale assessments, a much more mundane and more easily correctable factor affecting OTL is the scheduling of the assessment with respect to completion of instruction. Historically, most individual state assessment programs have had a prescribed testing window that was relatively narrow and consistent across the state, in the attempt to ensure that all schools had the same opportunity to instruct students prior to the assessment. When multistate assessment programs such as Smarter Balanced and PARCC were introduced, test administration guidelines were developed to maintain this same sense of equal opportunity to be exposed to instruction while accounting for differences in school year starting dates in states across the country. An example from the Smarter Balanced 2014–2015 technical report demonstrates how the percentage of a school's annual instructional days was used as a metric to minimize scheduling as an OTL factor that might affect the comparability of the test scores of individual students completing the Smarter Balanced assessment in different states (Smarter Balanced, 2016, p. 6-2) (see Box 2-2).

#### **BOX 2-2** **Test Administration**

Students in Smarter Balanced member states participated in the 2015 test administration once a specified percentage of the school year had occurred. Each state established a schedule for the administration of the Smarter Balanced summative assessments using a testing window as defined below:

##### **Grades 3–8**

- Testing shall not begin until at least sixty-six percent (66%) of a school's annual instructional days have been completed, and
- Testing may continue up to and including the last day of school.

##### **Grade 11**

- Testing shall not begin until at least eighty percent (80%) of a school's annual instructional days have been completed, and
- Testing may continue up to and including the last day of school.

States were allowed to establish more specific windows within the constraints of the consortium-defined windows described above.

SOURCE: Smarter Balanced (2014c, p. 25).

Given the connection between OTL and student achievement, it is no surprise that OTL is closely linked to large-scale assessments, specifically to the interpretation and use of results from large-scale assessments for high-stakes decisions (Pullin & Haertel, 2008). The Joint Standards define OTL in the context of testing as "the extent to which individuals have had exposure to instruction or knowledge that affords them the opportunity to learn the content and skills targeted by the test." The Joint Standards further state that opportunity to learn "has several implications for the fair and valid interpretation of test scores for their intended uses." The two testing

standards related to OTL directly address the interpretation and use of test scores for high-stakes decisions:

- **Standard 12.8.** When test results contribute substantially to decisions about student promotion or graduation, evidence should be provided that students have had an opportunity to learn the content and skills measured by the test.
- **Standard 3.19.** In settings where the same authority is responsible for both provision of curriculum and high-stakes decisions based on testing of examinees' curriculum mastery, examinees should not suffer permanent negative consequences if evidence indicates that they have not had the opportunity to learn the test content.

Although the connection among OTL, student achievement, and test scores is quite clear, the relationship between OTL and comparability is less straightforward. If a student has not had an opportunity to learn the content and skills measured by the test, the Joint Standards are clear that it would be inappropriate to use the student's performance on the test as the basis for a high-stakes decision such as promotion or high school graduation.

There is clearly a fairness issue with regard to the use of the individual student's test score. Judgment of the comparability of the student's test score may change, however, based on the assertions being made about student performance and/or the intended interpretation of student performance on the test. If the assertion is that the test score describes the student's current level of achievement in the content area then it is likely that the score is an accurate reflection of student achievement and could be considered comparable to the scores of other students taking the test.

Different assertions, however, could lead to different conclusions about comparability. Consider our original assertion, that the test score describes the student's current level of achievement in the content area, in relation to three alternate or additional assertions about student performance:

1. The test score reflects the achievement of students after they have received instruction in the content area being tested.
2. The test score reflects the level of achievement students can attain if they have had an adequate opportunity to learn the material.
3. The test score reflects what students could achieve at the next grade level or in college if provided an adequate opportunity to learn.

For each of these assertions, it would not be appropriate to consider the student's score comparable to the scores of other students who had the opportunity to learn the material.

In considering the original assertion and three alternative assertions presented above, nothing about the student's test score and its reflection of student achievement has changed. Our interpretation of the comparability of the test score due to OTL, however, did change, in that it is dependent upon the assertion being made or the expected interpretation of the score.

### **Familiarity with Item Formats and Tools Used on the Assessment**

An issue often conflated with a student's OTL is a student's opportunity to become familiar and comfortable with the item formats and tools that they will encounter on an assessment. Unlike OTL, however, where the impact on comparability is conditional on the claims and use of the test score, the lack of an adequate opportunity to become familiar with the item formats and tools used on the assessment is almost always a serious threat to comparability.

The distinction between OTL and familiarity with the assessment has to do with the expected relationship between a student's level of knowledge and skills, or achievement, and the score that the student will earn on the test. It is expected that two students with different levels of knowledge and skills due to differences in OTL will earn different scores on the assessment, that is, they will earn scores that accurately reflect their current level of achievement. However, if two students have equal levels of content knowledge and skills but differ in their familiarity with the item formats and tools used in the assessment, it is likely that the student who is more familiar and comfortable with the assessment will earn a higher test score. The observed difference in student performance would be attributed to construct-irrelevant variance (i.e., familiarity with the assessment) and the test scores would not be considered comparable.

The Joint Standards make it clear in several places that students should be provided an opportunity to become familiar and comfortable with the item formats that will be included on the assessment and also the tools, accommodations, and other supports that will be available or required for use during the test:

- **Standard 4.16.** The instructions presented to test takers should contain sufficient detail so that test takers can respond to a task in the manner that the test developer intended. When appropriate, sample materials, practice or sample questions, criteria for scoring, and a representative item identified with each item format or major area in the test's classification or domain should be provided to the test takers prior to the administration of the test, or should be included in the testing material as part of the standard administration instructions.
- **Standard 6.5.** Test takers should be provided appropriate instructions, practice, and other support necessary to reduce construct-irrelevant variance.
- **Standard 8.2.** Test takers should be provided in advance with as much information about the test, the testing process, the intended test use, test scoring criteria, testing policy, availability of accommodations, and confidentiality protection as is consistent with obtaining valid responses and making appropriate interpretations of test scores.

The recent transitions from paper-based to computer-based testing and from fixed-form to adaptive testing have introduced several examples of potential threats to comparability if appropriate steps are not taken in advance to ensure that students are familiar and comfortable with the item formats and tools they will encounter on the assessment. Examples follow of areas in which issues that are a potential threat to score comparability resulting from newly implemented computer-based tests have been encountered:

- **Familiarity with the test platform.** As a starting point, students must be comfortable working within an item and navigating through a test form. This includes comfort with the use of elements such as mouseover hover boxes, pop-up windows, or hyperlinks. Students must be familiar with the procedures for moving from one item to the next, skipping items (e.g., if an answer is required before moving on), and returning to previous items.
- **Familiarity with the student response device.** Students must be familiar with the response device, including the issues associated with the size of the screen and comfort with the use of a touch-screen, keyboard, or mouse, as required by the particular device.
- **Physical demands of responding to new item types.** In addition to the cognitive complexity of items associated with the new item types (e.g., multiple-select selected-response items, and technology-enhanced items such as drag-and-drop and hot spot), students also must possess the dexterity required to respond to the item.
- **Space available for written responses.** On a paper-based form students were provided with a fixed space to produce written responses (e.g., one page for constructed-response items and four pages for an essay). The computer-based version of the test used an expanding response box, meaning that students were given no visual cues about the expected length of a response.
- **Use of tools to respond to items on the mathematics assessment.** Students are required to use equation editors, graphing tools, and built-in calculators to respond on screen to mathematics items.
- **Presentation of reading passages and other stimuli.** Students are unable to view the test item and the reading passage at the same time.
- **Impact on common test-taking strategies.** When taking adaptive tests, students cannot apply strategies that they have been taught for other testing formats, such as previewing all of the items in advance and focusing first on items they can answer.
- **Familiarity with scoring rules for new item types.** When responding to multiple-select selected-response items, students may not be aware that selecting too few or too many options will result in the response being scored as totally incorrect.

Each of the potential threats described above can be mitigated by providing adequate opportunity for students to become familiar with the requirements of the computer-based test.

### Differences in Intended Uses of Test Results

Much like the case with OTL, it is widely accepted that differences in the intended uses of test results can affect the performance of individual students. Students taking a test that will be used for high-stakes decisions such as promotion to the next grade, high school graduation, or eligibility for a scholarship may perform differently than students who are taking the same test without such stakes attached to the results (Steedle & Grochowalski, 2017; Wise & Demars, 2005). Also similar to the case with OTL, the threat to comparability lies primarily in the interpretation of the student's test score.

The threat to comparability is greatest in cases in which there is a difference between the intended uses for the test when achievement level cut scores on the test are set and the intended uses when the test is administered. This would be the case within a state when achievement level cut scores are determined under one condition and then applied under a different condition. This would also be true across states when achievement level cut scores that are intended to be common across states are applied under low-stakes conditions in one state and high-stakes conditions in another.

The comparability of individual student scores is also threatened by the fact that it is virtually impossible to quantify and isolate the impact that content-based performance standards, intended uses, and outcomes (i.e., impact data) have on the process of identifying achievement level cut scores (i.e., standard-setting processes) for large-scale assessments.

As a concrete example of the threats to comparability described in this section, consider the common case of a judgment-based standard-setting process used to establish achievement level cut scores on a high school mathematics test. The central question asked of panelists during standard setting is some variation of the following: "Would a borderline-proficient student answer this item correctly?" A standard-setting panelist's response to that question is likely to be affected by the intended use of the test scores. Panelists will consider how likely it is that a student will persevere on an item that is complex, requires multiple steps to complete, or requires a written explanation to support a response. It is generally accepted by standard-setting panelists that students will be more motivated to persevere on such items when high stakes such as high school graduation are associated with performance on a test. Therefore, for a particular item, panelists might conclude that a borderline-proficient student is likely to answer the item correctly under high-stakes conditions, but unlikely to answer the item correctly if there are no student consequences attached to the test score.

When a decision is based solely on content and student motivation, standard-setting panelists are likely to set a higher achievement level cut score on a test that is used for high school graduation than on a test used for school accountability but with no stakes for students. However, panelists' overall judgments may be more influenced by student impact data on a high-stakes graduation test than on a test used for school accountability, resulting in a lower achievement level cut score on the high-stakes graduation test.

As stated above, because it is virtually impossible to isolate and quantify the various factors that might influence the location of an achievement level cut score, it is best practice to exercise caution when comparing scores on tests with different intended uses and stakes for students.

### **Differences Among Assessment Contractors**

By design or through the peculiarities of the procurement process, it is often the case that the "same" large-scale assessment is administered by different assessment contractors within or across years and/or within or across states. When the goal is to produce comparable test scores across assessment contractors, there are threats to comparability that might be hidden beneath the surface, which may affect score comparability even when it appears that all of the same procedures are being applied by each contractor: test delivery, scoring, and psychometrics. In each case, additional layers of specifications and tests to confirm outputs and outcomes may be necessary before declaring scores comparable.



### *Test Delivery*

Different assessment contractors are likely to employ different test delivery platforms, whether they are using their own proprietary platform or an open-source platform. When test forms are administered on different platforms within or across years (by the same or different assessment contractors), it is critical to confirm that the test platforms function in the same manner or to understand differences in how they function. Even with test forms administered on a single test platform, it is critical to confirm that the test platform functions in the same manner across local networks and allowable response devices. Answers to the following questions may affect student performance on the test and, therefore, affect score comparability (Way, Davis, Keng, & Strain-Seymour, 2016):

- Do test items and directions render the same way on each platform?
- Are the procedures students must follow to navigate through the test similar across platforms?
- Are the tools and supports provided to students as accommodations and/or required for use to respond to questions similar across platforms and accessed in equivalent ways?
- Is the speed with which items load and responses are submitted consistent across platforms?

### *Scoring*

When student responses are scored by two or more scoring contractors, it is critical to monitor scoring and ensure that scoring is consistent across contractors. This may seem obvious for items that are human scored, but it is also necessary for items that are machine scored. It is true even for multiple-choice items with a single correct response, which seem very easy to score consistently across contractors.

In large-scale testing programs that are still paper based, differences in how student responses are read and processed may lead to differences in how a student response is scored. Dependent on the specifications provided, settings applied, and equipment and materials used, marks that are recorded as a response or erasure by one system may be treated as a blank by another system. Such differences could result not only in different item scores, but also in different flags being applied to students or test administrators through data forensics procedures.

With regard to human-scored items, it is assumed that assessment contractors will apply the same scoring rubrics, use the same training materials, and attempt to apply the same training procedures (CCSSO & ATP, 2010). It may also be assumed that contractors will apply the same rules for recruiting and qualifying scorers and monitoring scorer consistency and accuracy throughout the scoring process. With all of those safeguards in place, score comparability could still be affected by differences in how scoring contractors arbitrate and resolve score differences or by the thresholds set for when to rescore items from a scorer flagged for scorer drift or inconsistent scoring.

When student responses to constructed-response items are scored by automated scoring engines, it is critical to verify to the extent possible that a student response will be scored the same way regardless of the scoring engine and scoring algorithm used

if the goal is to produce comparable individual test scores. In such situations, it is as important to verify that there is consistency in student responses to which an automated system will not assign a score as well as to verify consistency when scores are assigned.

### *Psychometrics*

When psychometric analyses such as item calibration, equating, scoring, and scaling are conducted by different assessment contractors, as with item scoring described in the previous section, it is assumed that certain procedures will be held constant (i.e., the same IRT model applied, and the same rules and procedures used for equating). To a greater extent than with test delivery and scoring, the software used and decisions made during psychometric analyses can be a threat to score comparability.

Across IRT software packages, there may be different ways the same IRT models are executed, producing differences in results for some individual students. Within and across software packages there are also many decisions that must be made during item calibration and scoring that could affect the comparability of results. There may also be differences across assessment contractors in the procedures for selecting samples of responses to use for item calibration. It is likely impossible to develop rules for all of the decisions that must be made during psychometric analyses. Best practice, however, requires a demonstration that psychometric analyses applied to the same set of student responses produce comparable results.

## CONCLUSION

We began with the assertion that most often the desired level of comparability when considering individual student scores on the same test is that test scores are interchangeable. On K–12 state assessments in particular, there is an expectation and assumption that if individual students had received a different form of the same test they would have received the same test score (within measurement error) and that the individual scores of two students taking the same test can be compared to the same achievement standard or to each other.

Next, we established that the term “same test” refers to a wide variety of cases in which students are taking tests involving different sets of items; those students have access to a range of supports, tools, or test variations, as needed; and the students may be taking the test under different modes of administration with different response devices. In fact, in practice, the least likely situation to be encountered would be one that fits the traditional perception of standardized testing or the colloquial definition of the term “same test”—students taking the same set of items under the same testing conditions.

Based on those parameters, we discussed approaches to achieving the desired level of comparability of individual test scores and threats to achieving that level of comparability. Ultimately, however, it will be necessary for a test developer to provide evidence to support a claim of test score comparability and for a test user (e.g., policy maker, school administrator, teacher, parent, or student) to be able to evaluate and accept or reject that claim. As with most things related to educational measurement and large-scale assessment there is not a single test to determine test score comparability

nor is there a simple yes-or-no determination of test score comparability that applies to all situations. The process of establishing the comparability of individual student scores on the same test involves compiling sufficient evidence to support the claims and inferences that will be made about student performance based on those test scores.

Our starting point for evaluating the comparability of individual test scores on the same test is differences. There may be differences in test items, administration conditions, scoring procedures, students' opportunity to learn the content or become familiar with the test itself, and/or the intended uses of the test results. Determining whether there is sufficient test score comparability involves determining the degree to which those differences individually or cumulatively affect student performance or the interpretation of student performance.

We discussed design and psychometric approaches to producing comparable test scores or, at least, minimizing threats to comparability. There are standards, guidelines, and established best practices in large-scale assessments to increase the likelihood that test scores will be comparable. At times, the definition of best practices requires trade-offs, compromise, and an uneasy coexistence of measurement principles and policy priorities or government mandates. A combination of qualitative judgments and quantitative analyses is needed to evaluate both the application of best practices and their impact in producing comparable test scores across test forms.

Informed, expert qualitative judgment, for example, may be sufficient to approve changes to a test design such as reducing the test length by one point, one passage, or one performance task. Those initial judgments, however, must be supported by quantitative analyses of the impact of the proposed changes on the reliability of the assessment, measurement error associated with individual students' scores, or the accuracy and consistency of achievement level classifications. Similarly, statistical analyses may indicate that alternative test forms have the same level of difficulty, reliability, and relationship with external variables. It is likely, however, that qualitative analyses are also needed to determine whether inferences from test scores and claims about student performance can still be supported.

Unfortunately, there is no simple answer to the question of whether individual students' scores on the same test are comparable. As concluded by the Committee on Equivalency and Linkage of Educational Test two decades ago, "Ultimately, policy makers and educators must take responsibility for determining the degree to which they can tolerate imprecision in testing and linking ... and responsible people may reach different conclusions about the minimally acceptable level of precision in linkages that are intended to serve various goals" (NRC, 1999, p. 4). In this chapter, we have attempted to provide the tools with which policy makers and educators can make informed decisions on the extent to which test forms and administration conditions have been designed to support the conclusion that individual students' test scores are comparable as well as the extent to which forces external to the test may affect inferences about the performance of particular students taking the test.

Within large-scale state assessment programs, standards and best practices have been identified to ensure that tests produce scores that are sufficiently comparable by tightly specifying factors such as test content, format, administration, scoring, and intended uses. Understanding the factors that enhance and threaten comparability and

evaluating those factors with regard to a particular test and inference about student performance on that test is the ongoing responsibility of all test users.

## REFERENCES

- AERA (American Educational Research Association), APA (American Psychological Association), & NCME (National Council on Measurement in Education). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.
- Banicky, L. A. (2000). *Opportunity to learn (Policy Brief Vol. 7)*. Retrieved August 1, 2019, from <http://udspace.udel.edu/bitstream/handle/19716/2446/opp+to+learn.pdf?sequence=>.
- Carroll, J. B. (1989). The Carroll model: A 25-year retrospective and prospective view. *Educational Researcher*, 18(1), 26–31.
- CCSSO (Council of Chief State School Officers) & ATP (The Association of Test Publishers). (2010). Scoring open-ended responses. In D. Eignor (Ed.), *Operational best practices* (Chapter 14). Washington, DC: CCSSO and ATP.
- Cooper, R., & Liou, D. D. (2007). The structure and culture of information pathways: Rethinking opportunity to learn in urban high schools during the ninth grade transition. *The High School Journal*, 91(1), 43–56.
- Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010). *Principles and practices of test score equating*. Princeton, NJ: Educational Testing Service.
- Friedlaender, D., & Darling-Hammond, L. (with Andree, A., Lewis-Charp, H., McCloskey, L., Richardson, N., et al.). (2007). *High schools for equity: Policy supports for student learning in communities of color*. Stanford, CA: School Redesign Network at Stanford University. Retrieved from [www.srnleads.org/resources/publications/pdf/hsfe/hsfe\\_report.pdf](http://www.srnleads.org/resources/publications/pdf/hsfe/hsfe_report.pdf).
- Haertel, E., Moss, P., Pullin, D., & Gee, J. (2008). Introduction. In P. Moss, D. Pullin, J. Gee, E. Haertel, & L. Young (Eds.), *Learning in doing: Social, cognitive and computational perspectives. Assessment, equity, and opportunity to learn* (pp. 1–16). Cambridge, UK: Cambridge University Press. <http://doi.org/10.1017/CBO9780511802157.003>.
- Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). Westport, CT: Praeger.
- Kurz, A. (2011). Access to what should be taught and will be tested: Students' opportunity to learn the intended curriculum. In S. N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), *The handbook of accessible achievement tests for all students: Bridging the gaps between research, practice, and policy*. New York: Springer.
- Lehr, C., & Thurlow, M. (2003). *Putting it all together: Including students with disabilities in assessment and accountability systems* (Policy Directions No. 16). Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved April 30, 2019, from <https://nceo.info/Resources/publications/OnlinePubs/Policy16.htm>.
- Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education*, 6(1), 83–102.
- MA DESE (Massachusetts Department of Elementary and Secondary Education). (2019). *Test administrator manual for computer-based testing, spring 2019*. Retrieved August 1, 2019, from <http://www.doe.mass.edu/mcas/testadmin/manual/TAM-CBT.pdf>.
- Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects* (Policy Information Rep.). Princeton, NJ: Educational Testing Service.
- NRC (National Research Council). (1999). *Uncommon measures: Equivalence and linkage among educational tests* (M. J. Feuer, P. W. Holland, B. F. Green, M. W. Bertenthal, & F. Cadelle Hemphill, Eds.). Committee on Equivalency and Linkage of Educational Tests. Board on Testing and Assessment. Commission on Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.
- PARCC (Partnership for Assessment of Readiness for College and Careers). (2017). *PARCC accessibility features and accommodations manual*. Retrieved April 30, 2019, from <https://parcc.pearson.com/manuals>.
- Pullin, D., & Haertel, E. (2008). Assessment through the lens of "opportunity to learn." In P. Moss, D. Pullin, J. Gee, E. Haertel, & L. Young (Eds.), *Learning in doing: Social, cognitive and computational perspectives. Assessment, equity, and opportunity to learn* (pp. 17–41). Cambridge, UK: Cambridge University Press. <http://doi.org/10.1017/CBO9780511802157.004>.
- Shepard, L., Hannaway, J., & Baker, E. (Eds.). (2009). *Education policy white paper on standards, assessments, and accountability*. Washington, DC: National Academy of Education.

- Smarter Balanced Assessment Consortium. (2016). Test administration. In *Smarter Balanced Assessment Consortium: 2014-15 technical report* (Chapter 6). Retrieved April 30, 2019, from <https://portal.smarterbalanced.org/library/en/2014-15-technical-report.pdf>.
- Smarter Balanced Assessment Consortium. (2018). *Smarter Balanced Assessment Consortium: Usability, accessibility, and accommodations guidelines*. Retrieved April 30, 2019, from <https://portal.smarterbalanced.org/library/en/usabilityaccessibility-and-accommodations-guidelines.pdf>.
- Steedle, J. T., & Grochowalski, J. (2017). The effect of stakes on accountability test scores and pass rates. *Educational Assessment*, 22(2), 111–123. <http://doi.org/10.1080/10627197.2017.1309276>.
- Way, W. D., Davis, L. L., Keng, L., & Strain-Seymour, E. (2016). From standardization to personalization. In F. Drasgow (Ed.), *Technology and testing*. New York: Routledge.
- Winter, P. C. (2010). Comparability and test variations. In P. C. Winter (Ed.), *Evaluating the comparability of scores from test variations*. Washington, DC: Council of Chief State School Officers.
- Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, 10(1), 1–17. [http://doi.org/10.1207/s15326977ea1001\\_1](http://doi.org/10.1207/s15326977ea1001_1).
- Zieky, M. (2014). An introduction to the use of evidence-centered design in test development. *Psicologia Educativa*, 20(2), 79–87.

# 3

## Comparability of Aggregated Group Scores on the “Same Test”

Leslie Keng and Scott Marion,  
*National Center for the Improvement of Educational Assessment*<sup>1</sup>

### CONTENTS

INTRODUCTION .....	50
Student-Level Versus Group-Level Comparability .....	50
Purposes and Uses .....	51
DERIVED SCORES .....	53
Measures of Central Tendency .....	54
Measures of Variability .....	54
Criterion-Based Measures .....	56
Growth and/or Value-Added Scores .....	57
Using Multiple Derived Scores .....	57
FACTORS AFFECTING COMPARABILITY OF AGGREGATE GROUP SCORES ...	58
Variations in Group Size and Composition .....	58
Variations Across Assessment Conditions .....	61
Variations in the Composition of the Assessment .....	62
Variations in Administration and Scoring Procedures .....	63
Differential Item and Test Functioning .....	63
PRACTICAL CONSIDERATIONS .....	64
Applying the Framework to Aggregated Score Comparability .....	65
An Example .....	70
CONCLUSION .....	72
REFERENCES .....	72

---

<sup>1</sup> We acknowledge the contributions of Susan Lyons to help us conceptualize this chapter and for producing Figure 3-2.



## INTRODUCTION

Chapter 2 outlined the challenges and opportunities associated with comparing scores among individual test takers on tests that are considered the same, such as the same end-of-year state achievement test. That discussion established key principles associated with comparing individual examinee scores. Some might think we must establish individual score comparability before establishing score comparability at various levels of aggregations such as school districts, states, and student groups, but there are many cases for which individual scores are not even generated (e.g., National Assessment of Educational Progress [NAEP], Programme for International Student Assessment [PISA]) and we still care deeply about aggregate comparability. Comparisons of aggregate scores, however, go beyond the typical units of analyses noted above and must include considerations of different test delivery platforms and modes of administration, the types of accommodations available to examinees in different settings, and many other factors. We limit this discussion to factors specific to group scores and do not rehash threats to individual score comparability discussed in the previous chapter.

Comparability is an essential requirement for establishing the validity of inferences of scores across individuals or other units, and validity is always evaluated in the context of specific purposes and uses. Therefore, after a brief introduction pointing out some differences between group and individual comparability is a discussion of the various uses and purposes associated with comparisons of aggregate performance for tests considered to be essentially the same. Following this initial framing, we describe the types of aggregate measures, or derived scores, used to compare group-level performance. We then present an analysis of some of the factors affecting the comparability of aggregate scores, drawing on examples from recent testing situations. This is followed by a discussion of comparability considerations unique to aggregate scores, such as when using matrix-sampling or computer-adaptive test designs. Because comparability exists on a continuum, there is rarely a dichotomous decision to indicate when comparability is either supported or violated. We conclude with a practical framework for evaluating and mitigating threats to the comparability of group scores.

### Student-Level Versus Group-Level Comparability

To illustrate how comparability of scores at the individual level does not guarantee that aggregates of the same scores are comparable, consider the example in Table 3-1 of a fictitious school (school A) across two academic years for the same test. Assuming all conditions for comparability of scores at the student level are met for this test (see Chapter 2), can we reasonably compare school A's performance across the 2 years and conclude that it has made significant improvements in eighth grade reading? Before the school starts celebrating its success in year 2, we should consider the notable drop in the number

**TABLE 3-1** School A's Performance on Grade 8 Reading Test Across Academic Years

	Year 1	Year 2
Number of students	120	50
Average score	425	512
Proficient	65%	80%

of students completing the test in year 2. This seems to indicate that there is something characteristically different between the composition of school A’s test-taking populations in years 1 and 2. Why has there been such a precipitous drop in the number of test takers?

Suppose further that we discover the important information in Table 3-2 about school A’s demographic composition for the eighth grade test takers in each year.

**TABLE 3-2** Demographics of School A’s Grade 8 Reading Test Takers Across Academic Years

	Year 1	Year 2
Free and reduced priced lunch	61%	30%
Special education	24%	12%
English learners	12%	4%

How comparable are the school’s average scores and percentage proficiency across the 2 years in light of this demographic shift? Is the improvement in year 2 due to the new academic initiative, or simply because of decreases in the participation of traditionally lower-performing student groups?

This example illustrates two factors that can affect the comparability of aggregated group scores even if we can assume the comparability of individual scores: group size and group composition. Before launching into a comprehensive discussion of these and other factors, we first discuss the importance of specifying purposes and uses associated with aggregated group scores.

### Purposes and Uses

It is axiomatic to say that validity is contingent upon intended purposes and uses and the claims users want to make based on the test scores. Given the close relationship between comparability and validity, it is fair to extend this axiom to comparability. Aggregated group scores from large-scale assessments are used for many purposes but generally fall into four major categories:

1. Monitoring population trends and patterns;
2. Comparing subgroup performance at specific time points and over time;
3. Evaluation of curriculum, instruction, interventions, and other programs; and
4. Accountability at various levels of the system (e.g., teacher, school, and district).

We expand on these broad purposes below and describe why comparability is essential to each of the categories, but to varying degrees.

### *Monitoring Full Population Trends and Patterns*

The National Research Council’s Committee on Developing Assessments of Science Proficiency in K–12 described monitoring as the most important function of large-scale assessments, especially as it pertains to the role of large-scale assessments in systems of assessment (NRC, 2014). Being able to accurately and reliably document academic performance and progress is critical to understanding how educational programs are

working and whether investments in education, implementation of new standards or curricula, or other major policies are contributing to large-scale improvements in educational systems. The information provided from such monitoring assessments often supports useful descriptive purposes.

NAEP, in operation for more than 50 years, is the most well-known monitoring assessment in the United States. Monitoring full U.S. population trends over this time period has been of paramount importance to document the nation's academic progress. While the "state NAEP" has been receiving more attention since its inception, likely because it allows for state-to-state comparisons, the "long-term trend NAEP" is a critical function of the program because it allows policy makers and other education stakeholders to track trends over years and decades on how the nation's schools and students are performing at a point in time and compared to prior performance. This discussion of NAEP highlights an important consideration that might be lost in our discussions of how to evaluate and maintain comparability even when comparability is threatened. Difficult lessons have been learned over the years about maintaining comparability when tests and/or testing conditions have changed. The infamous 1986 NAEP reading anomaly occurred when changes introduced to the test led to unanticipated score drops. When describing the extensive analyses into the score drop, Beaton and Zwick (1990) emphasized, "When measuring change, do not change the measure" (p. 165). In other words, the first strategy for maintaining individual- and group-level comparability should be to avoid changing the assessment, population, conditions, and other factors.

### *Comparing Subgroup Performance at Specific Time Points and Over Time*

In addition to comparing full population trends, evaluating the performance of subgroups of students, particularly educationally disadvantaged subgroups, has been a key component of major equity initiatives in the United States since before the passage of the No Child Left Behind Act of 2001 (NCLB). Full population trends portray a particular picture of educational performance for a particular entity, but such a picture may be misleading if the performance of multiple subgroups differs from the full population results. Therefore, being able to accurately and consistently compare the performance of subgroups of students is critical for sustaining a meaningful equity agenda.

### *Evaluation of Curriculum, Instruction, Interventions, and Other Programs*

A key purpose of many large-scale assessments is to support program evaluation efforts of states, school districts, and other educational entities. School districts and states expend significant resources on a variety of educational materials and programs. Therefore, district and state leaders must exert their fiscal responsibility by evaluating the extent to which such programs are fulfilling the intended aims. Beyond the direct fiscal rationales for pursuing evaluations of programs and interventions, there is an opportunity cost associated with pursuing a less effective compared to a more effective educational program. For example, students cannot be taught using two different mathematics curriculum programs at once; if it turned out they were using the less effective curriculum, any loss of learning would be an opportunity lost. Therefore, states and districts must have the information necessary to evaluate educational programs and interventions. The quality and usefulness of evaluation studies are dependent on many factors, but high-quality data are critical. Test scores often serve as outcome data and

essentially all evaluation designs rest on assumptions of comparability of data across groups (e.g., control and treatment groups) and over time.

### *Accountability at Various Levels of the System*

Finally, school and more recently educator accountability systems designed to meet federal and many state mandates have been designed with the intended purpose of supporting an equity agenda. After all, the original Elementary and Secondary Education Act of 1965 was a key component of President Johnson's "War on Poverty." Being able to identify schools needing support to help students succeed and recognizing schools that can serve as models for others generally requires comparable data across units (e.g., teachers, schools, and districts) to support the intended uses. Furthermore, many accountability systems include goals and targets based on changes in performance over time. Without assurances of comparability at both the individual and aggregate levels, such performance goals and targets are meaningless. Being able to support assumptions of comparability within the accountability system is critical to the credibility of accountability determinations resulting from the system. Generally, assessment results represent an important part of school and educator accountability systems, and the comparability of the assessment results at the aggregate level is often a necessary condition for ensuring the comparability of the full accountability system. Comparability is so important to accountability systems that states have performed some impressive statistical gymnastics to attempt to maintain comparability of the accountability system, but it is always much easier to defend inferences of comparability when there is evidence the assessment results are comparable.

Addressing the four main purposes for aggregate score comparability described above does not mean the results can be used to support causal claims, even though many policy makers would like to do so. Establishing causality requires well-thought-out designs controlling for variables and factors that can influence the results such as context effects (e.g., community characteristics, available resources, and school and district size) and educational variables (e.g., teacher and leader expertise and experience, educator turnover, student turnover, class size, and curriculum choices). Making an inference about educational effectiveness or other quality attributes based on test scores alone—whether measures at a single point in time or growth measures—often ignores these other factors that can operate as intervening variables to affect and/or explain the observed score patterns.

## DERIVED SCORES

The focus of comparability claims for individuals usually involves either student scores (e.g., raw score or scale score) or performance-level classifications on the same test. Group-level comparability considerations often involve *derived scores*, that is, measures that are a summary or aggregation of individual scores or classifications within the group. Derived scores are helpful because they help reduce large quantities of individual scores into a singular value, or statistic, that represents an important quantitative characteristic of the group. This makes comparisons at the group level easier for score users to manage and interpret. In general, derived scores can be categorized in four

general ways: measures of central tendency, measures of variability, criterion-based measures, and growth/value-added scores.

### Measures of Central Tendency

Most individual student reports include measures of central tendency, such as mean or median scale scores for the school, district, and state, to help provide context for the student's performance. Aggregate-level reports usually include mean or median scale scores that facilitate the comparison of student groups or entities across a state. Most school accountability systems use measures such as mean scale scores or mean and median growth scores as the basis of their indicators. Finally, measures of central tendency for groups or entities across time are used to establish trends and compare longitudinal performance.

While it is tempting to simply compare average performance across time, we must attend to things that could influence our interpretations or inferences about a change in the average score. For example, if a school's mean scale score on a test changed from 262 in the previous year to 275 in the current year, we might infer that the school's performance improved. There is certainly a higher score associated with the school now compared to previously. However, what if the population of the school changed substantially due to a shift in attendance boundaries? Or, what if the state made a change in the test, such as the removal of a traditionally more difficult writing task, that led to an unexpected increase in scores? In both (and other) cases, we need to exercise caution when making inferences across time or contexts.

Population or sample size, often referred to as  $n$ -count or simply " $N$ ," is an important consideration when computing measures of central tendency because it affects the implicit weight associated with each test score. For example, entities (e.g., schools) could have different multiyear averages depending on if they used weighted or unweighted approaches to compute the multiyear average. Two approaches could yield different 3-year averages, especially if the  $n$ -counts fluctuate significantly across years. If the  $n$ -count in year 1 is much smaller than that in the other 2 years, then test scores in year 1 would carry higher implicit weights in the average scale score calculation under an unweighted than under a weighted approach. The same  $n$ -count/implicit weighting consideration applies when we compute measures of central tendency across groups or entities of differing sizes and, in fact, can lead to issues of Simpson's paradox (see discussion below in the section "Implications: Simpson's Paradox").

### Measures of Variability

Measures of variability indicate the degree of spread or dispersion in a set of test scores. Common measures of variability include the range, interquartile range, variance, and standard deviation. When summarizing and reporting aggregate-level scores for a test, measures of variability are often overlooked or even omitted. This is likely because variability is in general less understood than measures of central tendencies. Even those who know the definitions of measures of variability may not appreciate the utility of these measures when comparing groups. It is important, however, to include measures

of variability in reporting and use them to aid in the interpretation and comparison of test results, at both the individual and aggregate levels.

To illustrate this, suppose a student achieves a score of 85 on a test and the mean test score for the class is 80. Our view of the student's performance would be different if the standard deviation of test scores for the class is 15 compared to if it is 2. In the former case, the student's score is certainly above the mean, but in the latter case, the student's score is likely one of the top scores. Without an understanding of variability and distributions, users would not make this inference.

At the aggregate level, consider the example of a weighted composite score that is used to determine a school's annual summative rating for accountability purposes. The composite score is calculated by applying a weight to each indicator in the accountability system. Consider the following hypothetical equation for computing a composite score that includes four accountability indicators: academic achievement (ACH), academic progress (PROG), English language proficiency (ELP), and chronic absenteeism (CA):

$$\text{Composite Score} = \text{ACH} \times 40\% + \text{PROG} \times 30\% + \text{ELP} \times 20\% + \text{CA} \times 10\%$$

The weights in this equation are referred to as *nominal* or *policy weights* because they are usually set to reflect policy priorities for each indicator in the accountability system. The equation above, for example, would serve to communicate a high-priority emphasis on academic achievement (by weighing it at 40 percent in the composite score), followed by academic progress (with a 30 percent weight). Note that this prioritization plays out in the computation of the composite score for a *given school*. Academic achievement, for example, accounts for 40 percent of the school's composite score and directly influences the summative rating associated with the score. However, if the primary interest is to *compare* schools on their composite scores, then the indicator that is most consequential *may not* be the one that has the highest policy weight. Consider the simple scenario in Table 3-3 of six schools and their composite scores, computed using the formula above, along with the standard deviation of each indicator score across the six schools.

**TABLE 3-3** Composite Scores for Six Hypothetical Schools

School	ACH (40%)	PROG (30%)	ELP (20%)	CA (10%)	Composite Score
A	60	50	57	97	60
B	61	51	21	93	53
C	59	50	95	92	67
D	60	51	45	90	57
E	61	49	82	92	65
F	59	49	37	91	55
Standard deviation <sup>a</sup>	0.9	0.9	27.9	2.4	5.6

<sup>a</sup> This is the standard deviation of each indicator across the six schools.

NOTE: ACH = academic achievement; CA = chronic absenteeism; ELP = English language proficiency; PROG = academic progress.



In this scenario, the ELP indicator has significantly higher variability, as indicated by the standard deviation values in the final row, than the other indicators. As a result, ELP becomes more consequential in distinguishing schools on their composite scores than the other indicators, even though it has a lower policy weight than ACH and PROG. This illustrates the general idea of *effective weights*, which is directly related to the degree of dispersion in the set of scores (as well as the policy weights) for each indicator or element in a composite score. The concept of effective weights in multivariate indicator systems, such as current school accountability systems, is important to our discussion of comparability. When policy makers establish nominal or policy weights, they believe they are establishing the metrics by which schools or other entities will be compared. However, the effective weights change the means of comparison so it is important for users to understand how the differences between nominal and effective weights can influence aggregate comparisons.

Both of these examples show the importance of considering measures of variability, in conjunction with measures of central tendency and criterion-based measures, to aid in the interpretation and comparison of individual and aggregated group scores.

### Criterion-Based Measures

Criterion-based measures are calculated based on how a group of scores compares to a criterion, such as a benchmark or standard. These measures are often expressed as proportions or percentages and are referred to as *rates* or *percent above cut* (PAC) measures. Common examples include proficiency rates (the proportion of scores that meet the cut score for “proficient” on a test), graduation rates (the proportion of students who have met the requirements for graduation), and chronic absenteeism rates (the proportion of students who meet the definition of “chronically absent”). A collection of related criterion-based measures can be used to facilitate more in-depth comparisons of aggregated group performance. For example, the proportion of students that would be in each performance level for various demographic student groups based on a set of panel-recommended cut scores is typically used as part of a standard-setting workshop to evaluate the reasonableness of the recommendations.

Criterion-based measures are often transformed mathematically into whole-number values and referred to as *indices* or *scores*. For example, proficiency rates are transformed from a percentage (e.g., 72%) to a whole-number score (e.g., 72) so that they can be combined with other related measures. The example in Table 3-3 (in the previous section) includes several accountability indicator values that are rates transformed to the 0-to-100 scale so that they can be combined into a composite score for each school.

Criterion-based measures, such as PAC, have appeal over measures of central tendencies and variability because they are thought to be more easily understood. For example, it appears more intuitively understandable to learn that school A’s “pass rate” on the test is 78 percent compared to reading that the school’s average scale score is 725 with a standard deviation of 15. In fact, states are required by federal law (currently the Every Student Succeeds Act) to report the percentage of students scoring at the proficient level or higher. It is also often the metric by which trend or gap measures, such as the “achievement gap” between student groups, is quantified. However, Ho (2008)

noted that PAC measures offer very limited and potentially misleading representations of group-to-group or longitudinal comparisons. For example, the relationship between the location of the proficiency cut score and the distribution of test scores can have major effects on apparent changes in trends. Table 3-4 illustrates a seemingly paradoxical case in which a school observes a substantial 15 percent increase in proficiency rate on a test in year 2 and a lesser increase of 5 percent in year 3. However, the change in average scale scores appears to tell a contradicting story.<sup>2</sup> Thus, the substantial jump in proficiency rate for year 2 was due primarily to the movement of students who were just below the cut (742) to above the cut. The improvement of students well below or well above the cut scores is not captured by the proficiency rates but is accounted for in the average scale score and standard deviation.

**TABLE 3-4** Derived Scores for a Hypothetical School Across 3 Years

Derived Score	Year 1	Year 2	Year 3
Proficient	45%	60%	65%
Average scale score	740	745	760
Standard deviation	20	18	15

### Growth and/or Value-Added Scores

Every grade testing under NCLB changed much in the U.S. testing context, but it also opened the door to documenting students' longitudinal performance. Two prominent approaches have emerged as the main methods for evaluating changes in student test scores over time: value-added modeling (VAM) (NRC, 2010) and student growth percentile (SGP) (Betebenner, 2009). Both VAM and SGP can have substantial effects on individual and group-level comparability, but discussions of VAM and SGP are beyond the scope of this chapter.

### Using Multiple Derived Scores

As we stressed and illustrated with examples in this section, when comparing the performance of groups on a given test, it is important to not limit the comparison to only one type of derived scores. One recommended practice is to take an initial look at the distribution of test scores in the groups of interest, via visual representations such as histograms, before even calculating any of the derived scores. Figure 3-1 shows the merits of visually inspecting the distribution test scores. In this simple example, both groups have the same mean and median scores, the same standard deviation, and the same proficiency rates on the same test (Test A). However, the histograms indicate that there is something characteristically distinct about the performance of the two groups, which could lead to different conclusions or have varying implications in terms of support or interventions for the groups.

<sup>2</sup> For example, if an effect size is computed using the difference in average scale scores and the (pooled) standard deviation, then it would show that the improvement in year 3 (from year 2) is more significant than that in year 2 (from year 1).

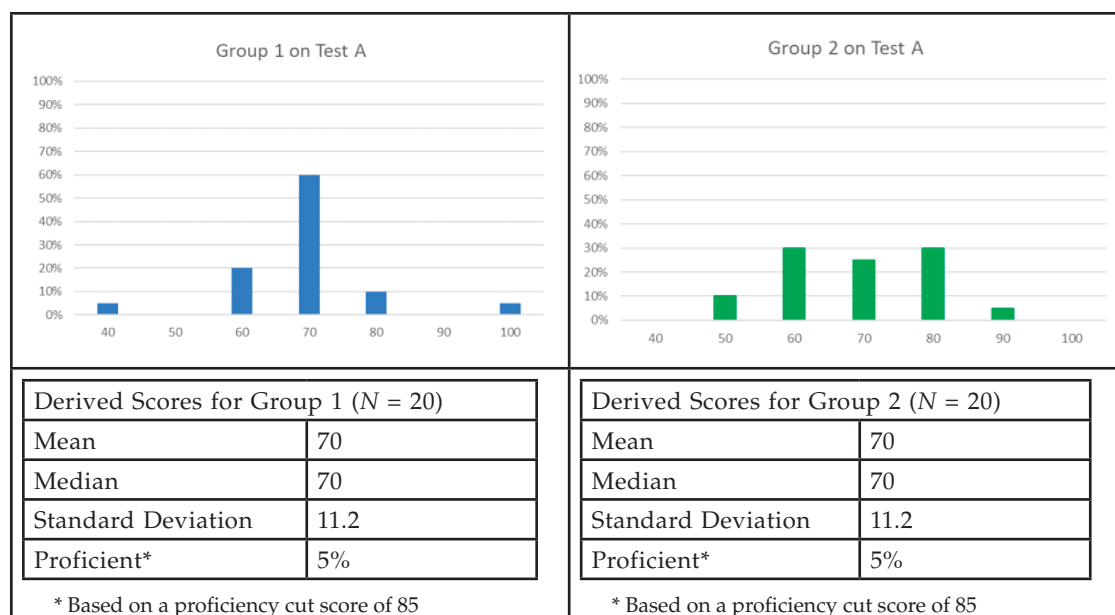


FIGURE 3-1 Visual comparisons of two groups of scores.

Even when we consider multiple derived scores and carefully examine the score distribution of two groups, there is still the essential question of whether the performance of two groups is comparable or, more generally, whether it is valid to compare two groups on their aggregated scores. To address this question, we need to take a close look at the contextual factors that can affect the comparability of derived scores.

### FACTORS AFFECTING COMPARABILITY OF AGGREGATE GROUP SCORES

In this section, we describe factors that affect the comparability of derived scores and the inferences that we can validly draw from comparing the aggregated group scores. We organize the factors into four broad categories: variations in group size and composition, variations across assessment conditions, variations in the composition of the assessment, and variations in administration and scoring procedures. Most of these factors also affect the comparability of individual student scores, as discussed in Chapter 2. In this section, we focus on the systematic issues influencing group-level score comparability.

#### Variations in Group Size and Composition

The initial example in this chapter illustrated how group size and composition can influence the comparability of aggregated group scores. In that example, we illustrated how these two factors raised questions about the comparability of the overall performance of a single school at two time points on the same test. In most applied scenarios, the ways group size and group composition interact and influence the comparability of aggregate group scores tend to be more complicated. We outline below several group

size and composition factors that could influence the comparability of aggregate scores over time and/or over jurisdictions.

### *Definition of Subgroup Across Jurisdictions*

The focus on equity and closing achievement gaps in educational systems, in addition to comparing the overall performance of groups, requires us to compare the performances of subgroups across jurisdictions, such as schools, districts, and states. In some cases, the size and composition of a subgroup with the same label may differ across jurisdictions because of the geographical factors associated with population distribution. For example, the “English learner” (EL) subgroup in a southwestern U.S. state, such as Arizona, New Mexico, or Texas is generally large (in both absolute size and percentage within the state) and consists mainly of students whose first language is Spanish. The EL subgroup for a northeastern state, such as Maine, New Hampshire, or Rhode Island, tends to be significantly smaller and comprises fewer students whose first language is Spanish, but instead has more students whose first language is Somali. The EL subgroup from Hawaii might be large but includes more students who are Asians or Pacific Islanders with a variety of first languages. The definition of subgroups across jurisdictions may also differ because of policy. For example, rules for entering and exiting EL status, for identifying students with disabilities, and for determining racial and ethnicity groups can vary across districts and states, leading to different sizes and composition of subgroups. Thus, a seemingly newsworthy headline such as “ELs in State A Significantly Outperformed ELs in State B on the SAT Math Test This Year” could be misleading. Instead of drawing conclusions about the effectiveness of academic interventions or support programs for ELs in state A (or lack thereof in state B), it would be prudent to first carefully examine the size and composition of the EL subgroups in each state. For more on the comparability challenges associated with evaluating and maintaining such comparability, see Chapter 6, *Comparability When Assessing English Learner Students*.

This issue must be considered for other subgroups as well and not just for comparisons across states. For example, special education rates are notoriously variable across states and across districts within states. Even if the proportion of special education students in the population remains steady over several years, “special education” is an amalgamation of 14 specific disabilities and the constellation of the proportion of students with these specific disabilities can vary considerably even if the total proportion of special education students does not change. A shift in the makeup of the special education subgroup, such as a noticeable increase or decrease in students with intellectual disabilities compared with speech or language impairment, can lead to measurable changes in the performance of the special education subgroup. For more on the comparability of assessments concerning students with disabilities, see Chapter 7, *Comparability When Assessing Individuals with Disabilities*.

A casual reader might think these issues are unique to “educational” subgroups, such as ELs and students with disabilities, and not related to “natural” or “socially defined” subgroups, such as racial, ethnic, or poverty-related subgroups, but that is not true. The challenges faced by economically disadvantaged students have been well documented, but many acknowledge the differences between rural and urban poverty or the differences between those just below the poverty line and those far below.

Similarly, the Hispanic, African American, and Asian/Pacific Islander student groups could all vary considerably in the makeup of each subgroup in ways related to both performance and culture. The main point here is that comparing the performance of both the total population and specific subgroups over time involves understanding how the proportions of student groups have varied over time and how the constellations of the smaller subgroups vary within the larger student groups.

### *Group Size and Sampling Error*

State accountability and assessment leaders have learned a lot about the effects of group size on sampling error. Statistical purists might bristle at the term “sampling error” because many contend that a group of students tested in a particular year is a population. These debates played out early in the NCLB era when states first had to determine the minimum number of students needed to constitute a subgroup (i.e., minimum  $n$ ). States were required by law to make valid and *reliable* determinations, but were also expected to include as many students and subgroups in the accountability determinations as possible. Researchers and state leaders witnessed the notable influence of group size on the variability of the estimates of indicator scores (e.g., percent of students scoring at the proficient level or graduation rates) and had to wrestle with the trade-offs between “reliability” and consequences associated with including as many subgroups as possible in accountability determinations (Kiplinger, 2008; Linn & Haug, 2002). While state leaders recognized the importance of reliable classifications, they quickly learned they would need minimum group sizes so large to meet reasonable reliability thresholds they would exclude many student groups from accountability. Even though many states used confidence intervals around score estimates for smaller groups (e.g., Marion et al., 2002), it became apparent that smaller groups had more volatile score trends than larger groups or schools. Therefore, group or school size is an important consideration for aggregate-level comparability because smaller schools (subgroups) bounce in and out of accountability determinations at higher rates than larger entities (e.g., Linn & Haug, 2002).

### *Changes Over Time Within Jurisdictions*

The size and composition of a group of students within a school, district, or state could change over time. Many schools and districts are in neighborhoods with highly transient populations. Natural disasters can have a significant impact on the constitution of jurisdictions at specific points in time. For example, Hurricane Katrina displaced millions of residents in the state of Louisiana in 2005, affecting the size and composition of school and districts not only in Louisiana, but also in its neighboring states in the Gulf Coast region of the United States. The recent gentrification within large U.S. cities has led to the movement of families with higher socioeconomic status (SES) to traditionally low-SES areas, changing the makeup of schools and districts in both urban and suburban neighborhoods.

Changing definitions or criteria for benchmarks or eligibility rules can also affect the group of test takers. For example, to better align with college and career benchmarks, the WIDA Consortium, which among other things develops assessments for ELP,

adjusted the cut scores of its ACCESS for ELLs 2.0 assessment starting in the 2016–2017 academic year. Many states use performance on ACCESS to determine whether an EL student has attained ELP. The number and composition of ELs who meet the ELP eligibility criteria are likely different in the years before and after the adjustment to the ACCESS cut scores.

Finally, politically motivated initiatives, such as the recent “opt-out” movement in several states where parents elect to excuse their children from taking standardized statewide assessments, can have a substantial effect on the size and composition of the test-taking population depending on the degree of opt-outs across the years in each jurisdiction. The opt-out movement likely had a bigger effect on comparability of statewide achievement test scores and on accountability results in several states with substantial opt-out rates, such as New York, Colorado, and Utah. However, even states with apparently minor opt-out issues can still face comparability challenges because students who opt out generally are a nonrandom portion of the tested population both within and across years.

### ***Implications: Simpson’s Paradox***

Simpson’s paradox is a well-known statistical phenomenon manifest in the social sciences when the underlying population (or sample) is composed of subgroups and comparisons are being made across time or occasions (Blyth, 1972). The issues related to subgroup definitions and compositions described above may play out as a Simpson’s paradox. This paradox gained notoriety with Wainer’s (1986) explanation of the SAT score increases in the early 1980s.

The average total SAT score increased by 7 points from 1980 to 1984, yet the average score for whites increased by 8 points and 15 points for nonwhites during this time frame. Given the score increases for whites and nonwhites, many wondered why the overall increase was not somewhere between 8 and 15 points. Wainer explained that because the nonwhite scores started so much lower, their score increase of 15 points was not enough to bring them up to the performance of the white scores or even the overall average. Therefore, the weighted average score increase takes into account the size of each group, their starting point, and their score increase. This example demonstrates why it is important to pay attention to the potential of Simpson’s paradox when making comparisons over time for an entity comprised of differentially performing subgroups.

### **Variations Across Assessment Conditions**

Chapter 2 discussed several threats to individual score comparability related to differences in assessment conditions. The threats include factors such as the mode of administration (e.g., paper versus desktop, or laptop versus tablet), test takers’ familiarity with item formats, accessibility features and accommodation tools, availability of software and/or hardware for computer-based testing, and the general environment or context in which the assessment is administered. The salient point is that variations across assessment conditions could influence comparisons across groups because these potentially confounding factors are nonrandomly distributed across entities.



For example, a student's performance on a test could be affected by their level of comfort with responding to novel or "innovative" item types, navigating the computer-based testing interface, and/or taking a test on the specific digital device available in the student's school or testing location. In this case, differences in student scores on the test could indicate their experience or level of exposure to computers as much as their math achievement. When these differences are aggregated to the group level, it can manifest as performance differences that are exacerbated by access to technology or the level of computer literacy within a school or district. In other words, differences between groups that are referred to as "achievement gaps" may be due, in part, to gaps in technology access and/or technology literacy.

Several testing programs in recent years have wrestled with issues of mode comparability as school districts and states migrated from paper- to computer-based administrations. These differences were related to several of the issues discussed above (e.g., familiarity) and created challenges for state leaders. On the one hand, they were eager to shift their testing programs online, but on the other, they were reluctant to disadvantage any schools or subgroups that had not yet become used to the new testing system. Depending on the degree of novelty, such as with technology-enhanced items, these effects tended to be observed in the elementary grades' language arts performance—often writing—and the effects would dissipate after a few years. However, for schools that were concerned about accountability results, waiting a few years was not a satisfactory option. Therefore, several states conducted mode comparability studies and proposed making adjustments to the scores associated with the lower-performing mode. While this sounds straightforward, it was not. Rarely are the effects of mode uniform across the score distribution, so a single mean adjustment would not address the problem as fairly as intended. Therefore, many state leaders opted for a policy response by offering a "hold harmless" for schools experiencing score declines consistent with the shift to computer-based testing. In this case, state leaders offered schools the option of using the results based on computer-based testing or maintaining their accountability levels from the last year of paper-based testing, whichever was higher. The policy option was generally offered for a single transition year.

### **Variations in the Composition of the Assessment**

There are cases in which states or districts may make changes to the design of an assessment so what is touted on the surface as the "same test" may not in fact be the same in composition. Both Common Core-based consortia have examples of such modifications to existing operational tests. One state in the Smarter Balanced Assessment Consortium decided to remove some of the English language arts (ELA) performance tasks from its test forms. New Meridian Corporation, who manages what was formerly known as the Partnership for Assessment of Readiness for College and Careers (PARCC) consortium, made a notable change when it started offering a shorter blueprint in 2019 that several member states chose to administer. In both the Smarter Balanced and New Meridian cases, two group-level score comparability issues should have been considered: the comparability of scores over time within the same jurisdiction and the comparability of scores at the same time across jurisdictions. For example,

the Smarter Balanced state mentioned above should have evaluated whether proficiency rates and aggregated scores, such as average scale scores and growth measures, reported before and after the removal of the performance tasks were comparable within the state and therefore appropriate for use in its accountability system. If the state found larger-than-expected differences compared to normal year-to-year fluctuations, it could have decided to restart its accountability trends or it could have tried to link the two sets of scores using equipercentile linking or some similar method. Smarter Balanced should have considered whether it would be reasonable to compare the state results with other member states after the adjustment. For example, if the state's score change after removing the performance tasks was noticeably different than that of other states, especially in terms of subgroup performance, the consortium could have considered eliminating this state from consortium average performance.

### **Variations in Administration and Scoring Procedures**

Even if the test design is identical and assessment conditions are controlled for to the extent practicable, different entities or jurisdictions may vary in their approaches to implementation and rigor in enforcing the administration policies and scoring procedures for the "same test." While these are factors that can influence the comparability of individual scores, it is particularly noteworthy for aggregated group scores when different testing vendors or contractors are responsible for administering and scoring the test across time or across states. Many of these factors, such as different test security protocols, scorer qualifications, and psychometric procedures, were discussed in Chapter 2, but we emphasize that such variations may negatively affect the comparability of group scores from the same test within a state or jurisdiction across time, or across states or jurisdictions at a given point in time.

### **Differential Item and Test Functioning**

The four types of testing variations just discussed can affect state, district, school, and subgroup comparability within and across years. This lack of comparability could play out similarly among subgroups, but often the threats function nonrandomly across subgroups. Differential item functioning (DIF) and differential test functioning (DTF) encompass a substantial set of conceptualizations and analytic techniques used to evaluate these nonrandom outcomes across subgroups and can help shed light on the effects of noncomparability on aggregate-level performance.

DIF is said to occur when two or more sets of examinees, who are otherwise of equal ability (achievement), perform differently on specific items (AERA, APA, & NCME, 2014). In other words, when examinees have the same total test score, there would be no reason to expect systematic performance differences on any item on that test. When such differences occur, typically beyond prespecified thresholds, the item is said to function differentially for particular subgroups of students. Evidence of DIF is not necessarily evidence of test bias (Camilli & Shepard, 1994). Because investigations of item or test bias seek to determine whether scores for subgroups of students may be affected by attributes other than those the test is intended to measure, DIF procedures

may help shed light on the degree with which variations due to assessment compositions, conditions, and administration and scoring processes contribute to differential performance across subgroups that is unrelated to the measurement target. If DIF is detected, content, bias, and assessment experts are convened to try to ascertain whether evidence of item bias exists.

One could imagine a scenario where a large set of items on a test exhibits slight DIF, but none of the items are flagged for meeting prespecified criterion values; however, the direction of the DIF is consistent (i.e., favoring the same group). This could be due to a test that is functioning differently for various subgroups of students. DTF is like DIF, but based on the total test form (AERA et al., 2014). In DIF, however, the total test score is used to contextualize item performance. We need to use a different criterion, obviously, to evaluate DTF, and scores or performance on related measures or other external criteria are used in evaluations of DTF.

Again, observations of DIF or DTF do not mean the test is biased against specific subgroups of students. DIF also may be an indicator of multidimensionality when the test is being treated as a single dimension. DIF and DTF require reasonably sized samples (e.g.,  $n = 200$ ) in order to conduct the analyses. Test makers are not off the hook with smaller samples because they can pursue qualitative approaches, such as cognitive laboratories, to investigate whether items are functioning as intended and similarly for various subgroups.

## PRACTICAL CONSIDERATIONS

Comparability is critically important for monitoring performance trends over time within and across groups; otherwise, educational leaders will not be able to accurately judge if their improvement efforts are working. Additionally, essentially all state accountability systems rely on strong assumptions of comparability to support normative comparisons (e.g., lowest-performing 5 percent of schools) and longitudinal comparisons (e.g., school progress toward long-term and interim goals). These assumptions require evidence documenting the threats to aggregate-level comparability and are not strong enough to invalidate the comparisons. In this section, we provide practical guidelines for practitioners and score users faced with the challenges of needing to make inferences and to act on conclusions drawn from imperfect group-level comparisons of assessment outcomes.

A popular adage in medicine is, “Prevention is better than cure.” We suggest that the same idea applies to supporting the comparability of test scores at both the individual and aggregate levels. That is, the optimal approach to supporting claims of comparability is not a series of post hoc analyses, but rather it should begin with the design of the assessment system itself. It means *planning* for factors that may be threats to comparability during the development of the assessment system, *evaluating* the degree to which the threats are mitigated as the system is implemented, and then, if necessary, *adjusting* for any manifested threats or differences. Most importantly, much thought and planning should be put into *communicating* with the field and end users about appropriate score comparisons and interpretations. Figure 3-2 is a visual representation of this framework with a key guiding question for each step.

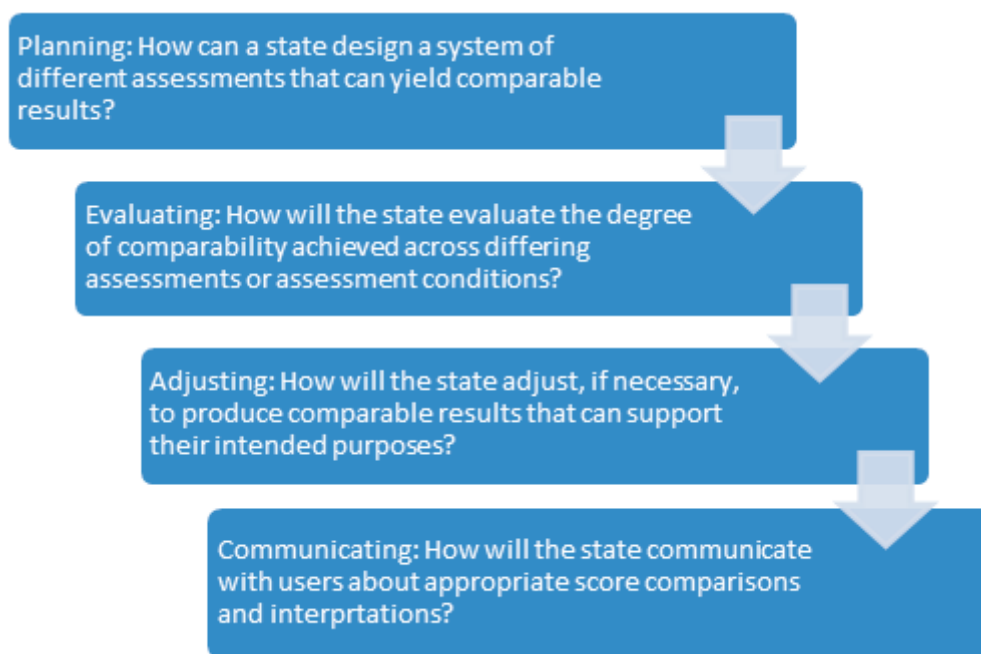


FIGURE 3-2 Framework for supporting comparability claims in a state assessment system.

The order of these guiding questions is very important. It would not be possible to evaluate the factors influencing group-level comparability if comparability has not been carefully considered in spite of these factors. No amount of evaluation and adjustment can fix a system that has not been carefully designed to produce comparable scores. Thus, garnering evidence to support comparability of the test results requires thoughtful planning of the processes that promote comparability, and program monitoring mechanisms for evaluating comparability. Additionally, states must have a clear plan for effectively communicating with the field about the degree to which scores can be meaningfully compared among groups and entities, over time and across assessment conditions.

States should also consider the *people* who can provide support in each step of the framework. When identifying these supporting parties, the state should think not only about assessment and accountability professionals within the state education agency, but also those in local districts and schools; staff from its testing vendor or subcontractors; established practitioners and experts in the field, such as those on the state’s various advisory committees; and educational stakeholders from the community, such as policy makers, teachers, and parents.

### Applying the Framework to Aggregated Score Comparability

To illustrate the use of this framework in the context of supporting the comparability of group-level scores, we provide key questions and considerations that a state can consider at each step to help mitigate or minimize the threats to comparability described earlier in the chapter. We recall the threats we focused on above:

- Variations in group size and composition,
- Variations across assessment conditions,
- Variations in the composition of the assessment, and
- Variations in administration and scoring procedures.

Table 3-5 below summarizes the key questions and/or considerations for each of the threats to group-level comparability for each of the steps in the framework.

**TABLE 3-5** Considerations for State Leaders to Help Mitigate or Minimize the Threats to Comparability

Comparability Threat	Variations in Group Size and Composition
<b>Planning</b>	<ul style="list-style-type: none"> <li>• What is the range of group sizes that the state observes within a given year? Across multiple years?</li> <li>• How similar or different are the students that make up the groups in the state within and across years in terms of key demographic and educational characteristics?</li> </ul>
<b>Evaluating</b>	<ul style="list-style-type: none"> <li>• Is there a minimum group size at which derived scores are no longer reliable?</li> <li>• What is the degree of uncertainty (e.g., standard error) of the aggregate scores for different group sizes?</li> <li>• To what extent are the student characteristics that vary across groups correlated to the group's performance?</li> </ul>
<b>Adjusting</b>	<ul style="list-style-type: none"> <li>• For each purpose and use that the state is comparing groups, is it reasonable to combine or collapse certain groups (e.g., form a "super-subgroup") to increase group sizes or make the groups more similar in size?</li> <li>• Are there statistical adjustments that that state can make to account for the larger degree of uncertainty associated with small group sizes?</li> </ul>
<b>Communicating</b>	<ul style="list-style-type: none"> <li>• If no adjustments are made, what explanations or disclaimers should the state include with the results of group comparisons to address the influence of uncertainty or precision resulting from differences in group sizes and/or composition?</li> <li>• If adjustments are made to account for the variation in group sizes and/or composition, what information should the state include with the results of group comparisons to explain the adjustment procedures and rationale as well as the uncertainty associated with such adjustments?</li> </ul>

TABLE 3-5 Continued

Comparability Threat	Variations Across Assessment Conditions
<b>Planning</b>	<ul style="list-style-type: none"> <li>• What protocols, instructions, and support can the state implement to minimize the impact of context effects due to variation in assessment conditions <i>across the general test-taking population</i>?</li> <li>• What protocols, instructions, and support can the state implement to minimize the differential impact that assessment conditions can have on <i>specific subgroups</i>?</li> </ul>
<b>Evaluating</b>	<ul style="list-style-type: none"> <li>• Are there any group-level performance trends that are correlated with specific assessment conditions?</li> <li>• Is there evidence of subgroups that are differentially affected by certain assessment conditions?</li> <li>• Does the impact of any assessment conditions on group-level performance change (i.e., either weaken or grow stronger) over time?</li> </ul>
<b>Adjusting</b>	<ul style="list-style-type: none"> <li>• Should the state apply statistical adjustments to account for any of the following: <ul style="list-style-type: none"> <li>◦ An overall (main) effect for an assessment condition (e.g., a “motivation” or “opportunity to learn” adjustment for all students)?</li> <li>◦ A differential (interaction) effect for an assessment condition and a subgroup of students (e.g., a “mode” adjustment for students who take the test online)?</li> <li>◦ A change in the effect of an assessment condition over time (e.g., a “familiarity” effect applied to group-level scores in subsequent years of an assessment program)?</li> </ul> </li> </ul>
<b>Communicating</b>	<ul style="list-style-type: none"> <li>• If no adjustments are made, what explanations or disclaimers should the state include with the results of group comparisons to address the potential impact of variations in assessment conditions?</li> <li>• If adjustments are made to account for the variations in assessment conditions, what information should the state include with the results of group comparisons to explain the adjustment procedures and rationale?</li> </ul>

continued



TABLE 3-5 Continued

Comparability Threat	Variations in the Composition of the Assessment
Planning	<ul style="list-style-type: none"> <li>• If changes in the composition of the assessment have been mandated, how can the state approach the changes to the test blueprints, design, content specifications, and/or performance-level descriptors, etc., to minimize the impact on comparability?</li> <li>• Can the state propose alternatives to changing the assessment composition or request longer timelines to implement the change?</li> </ul>
Evaluating	<ul style="list-style-type: none"> <li>• What impact does the change in assessment composition have on the underlying scale, performance standards (i.e., cut scores), reliability, and validity of the assessment?</li> <li>• Do the changes in assessment composition differentially affect certain groups of students in the state?</li> </ul>
Adjusting	<ul style="list-style-type: none"> <li>• Are the changes in assessment composition so substantial that the state cannot maintain the existing scale or cut scores?               <ul style="list-style-type: none"> <li>○ If so, what processes does the state need to implement to generate a new scale and cut scores?</li> <li>○ If not, what adjustments, if any, should be made to the existing scale or cut scores?</li> </ul> </li> </ul>
Communicating	<ul style="list-style-type: none"> <li>• What information should the state provide to the field about the changes to the assessment composition and any potential implications to group-level performance comparisons?</li> <li>• If a new reporting scale and cut scores are introduced, or the existing scale and cut scores are modified, what guidelines can the state provide to help the field interpret the assessment outcomes before and after the change? What cautions or disclaimers should the state provide in terms of interpreting group-level trends over time?</li> </ul>

TABLE 3-5 Continued

Comparability Threat	Variations in Administration and Scoring Procedures
<b>Planning</b>	<ul style="list-style-type: none"> <li>• What training, documentation, and real-time support can the state provide to local testing personnel to ensure that the administration procedures are implemented with fidelity?</li> <li>• What test security protocols and procedures does the state need to enforce to minimize testing irregularities or improprieties during administration?</li> <li>• What qualification criteria, scoring protocols, and monitoring procedures should the state put in place to support reliable scoring processes, including both machine and human scoring?</li> <li>• How can the state minimize the impact of transitioning to innovative scoring approaches on score comparability?</li> </ul>
<b>Evaluating</b>	<ul style="list-style-type: none"> <li>• What evidence does the state need to collect to confirm that administration procedures have been implemented with fidelity?</li> <li>• What data forensics analyses should the state conduct to detect potential testing irregularities or improprieties?</li> <li>• What metrics should the state calculate and monitor regularly to confirm that the scoring processes are reliable and implemented with fidelity?</li> <li>• Does the state have evidence of differential scorer effects on responses from different subgroups?</li> <li>• What research studies does the state need to conduct to support the validity of innovative scoring approaches?</li> </ul>
<b>Adjusting</b>	<ul style="list-style-type: none"> <li>• If there is evidence that administration or scoring procedures have not been implemented with fidelity, what adjustments, if any, does the state need to make to affected student scores? Does the state need to apply any adjustments to group-level scores?</li> <li>• If there is evidence of testing irregularities or improprieties, how should the state handle the student scores in question? Should the state apply any adjustments to group-level scores?</li> <li>• If there is evidence of differential scorer effects on specific subgroups, what adjustments should be made to student scores in the impacted groups? Should the state apply any adjustments to aggregated scores for the impacted groups?</li> </ul>
<b>Communicating</b>	<ul style="list-style-type: none"> <li>• If there are issues related to test administration, scoring, or incidents of testing irregularities or improprieties, how can the state communicate the issues, potential impacts, and mitigation strategies to the field in a clear and transparent manner?</li> </ul>

### An Example

Consider a scenario in which a state is legislatively required to remove writing from its ELA assessments. To minimize the potential impact on longitudinal trends, the state would like to maintain comparability of the ELA scale score and performance levels. How should the state approach this change? Table 3-6 outlines potential approaches the state may employ to evaluate and perhaps maintain comparability. Note that, based on our categorization of threats to group-level score comparability, the removal of writing from ELA is a variation in the composition of the assessment, and perhaps a significant variation.

**TABLE 3-6** Example Application of Comparability Support Framework

Framework Step	Potential Courses of Action
<b>Planning</b>	<ul style="list-style-type: none"> <li>• The state conducts comparative analyses of the old and new test blueprints, design, content specifications, and achievement-level descriptors to determine whether the underlying ELA construct is substantively affected by the removal of writing prompts.</li> <li>• The state convenes meetings with ELA content specialists and educators from across the state to provide input and feedback on the proposed changes to the test blueprints, design, content specifications, and performance-level descriptors.</li> <li>• The state examines its school accountability system and identifies aggregate measures, indicators, classifications, and/or identification business rules that are potentially affected by the removal of writing from the ELA assessments.</li> </ul> <p><i>Supporting Parties</i></p> <ul style="list-style-type: none"> <li>• ELA content specialists and educators from the state education agency, testing vendor, and representatives from across the state</li> <li>• Accountability specialists at the state education agency</li> </ul>

TABLE 3-6 Continued

Framework Step	Potential Courses of Action
<b>Evaluating</b>	<ul style="list-style-type: none"> <li>• The state conducts empirical studies, based on data from the most recent operational administration, to evaluate the impact of removing writing tasks on item calibration, scaling, test reliability, predictive validity, and classification accuracy and consistency for the ELA assessments. The studies are conducted at each grade level for all students and by subgroups.</li> <li>• The state replicates the empirical studies during the upcoming operational administrations.</li> <li>• The state continues to monitor for unexpected shifts in ELA performance, especially for the female student group (which traditionally scores higher on writing) and schools that previously showed notable improvement in writing.</li> <li>• The state performs impact analyses with its accountability system to evaluate whether there are any unexpected changes in school ratings or identifications because of the removal of writing. If the analyses reveal such changes, the state examines the affected schools to see if there are any discernable trends in terms of the characteristics of the schools. If the state judges the trends to be substantial, the state may choose to reset accountability goals and establish a new baseline.</li> </ul> <p><i>Supporting Parties</i></p> <ul style="list-style-type: none"> <li>• Psychometric and research experts</li> <li>• Technical advisory committee (TAC)</li> <li>• Accountability implementation specialists and programmers</li> </ul>
<b>Adjusting</b>	<ul style="list-style-type: none"> <li>• Based on the findings from the empirical analyses, the state makes adjustments to the underlying scale or establishes a new scale for its item bank.</li> <li>• The state convenes an ELA standards validation meeting to recommend potential adjustments to the cut scores on the ELA assessments. Depending on the committee's recommendations, the state adjusts its reporting scale.</li> <li>• If unexpected shifts in ELA performance are detected in subsequent years, the state considers additional adjustments to the scale and cut scores.</li> <li>• Based on changes made to the assessment system and the impact analysis on the accountability system, the state makes decisions such as whether to adjust the affected accountability components (i.e., aggregate measures, indicators, ratings, etc.), to introduce new accountability components, and/or to suspend reporting of accountability outcomes during the transition year.</li> </ul> <p><i>Supporting Parties</i></p> <ul style="list-style-type: none"> <li>• Psychometric and research experts and TAC</li> <li>• ELA content specialists and educators from across the state (to participate in the standards validation meeting)</li> <li>• Accountability implementation specialists and programmers</li> <li>• Accountability leadership and advisory committee</li> </ul>

*continued*

TABLE 3-6 Continued

Framework Step	Potential Courses of Action
Communicating	<ul style="list-style-type: none"> <li>• The state convenes focus and/or advisory groups to review and provide input on the updated individual student reports (ISRs).</li> <li>• The state makes changes to the ISRs based on feedback from the focus and/or advisory groups.</li> <li>• The state organizes community outreach meetings to explain the changes to the assessments, especially in terms of whether schools and districts can maintain trends or if new baselines must be established. In addition to clearly communicating the decisions, a key communications goal is getting buy-in from key stakeholder groups.</li> <li>• The state publishes communication resources that highlight findings from the empirical studies, outcomes from the standards validation process, changes to the reporting scales, and impacts to the accountability system components and outcomes.</li> <li>• The state updates its annual assessment and accountability technical manuals with details about the empirical studies, changes to the scale and cut scores, and changes to the accountability system components.</li> </ul> <p><i>Supporting Parties</i></p> <ul style="list-style-type: none"> <li>• Assessment and accountability reporting specialists</li> <li>• District and school administrators</li> <li>• Community leaders and educational stakeholders</li> <li>• Psychometric and research experts and TACs</li> </ul>

## CONCLUSION

The focus of this chapter has been on the comparability of aggregated group scores. In our experience, states often attend to the comparability of scores at the individual student level because they are perceived as having a more direct impact on students. Less attention is often afforded to score comparability at the aggregate level. We have attempted to highlight the importance of considering the comparability of group-level scores by describing the purposes and uses of comparing scores at the aggregate level, citing common threats to group-level score comparability, and proposing a framework that states can use to evaluate and build its case for comparability. Chapter 2 and this chapter should provide assessment and accountability professionals with broad knowledge and practical guidance on how to establish the validity of test scores and their inferences through comparability at all levels of reporting.

## REFERENCES

- AERA (American Educational Research Association), APA (American Psychological Association), & NCME (National Council on Measurement in Education). (2014). *Standards for educational and psychological measurement*. Washington, DC: AERA.
- Beaton, A. E., & Zwick, R. (1990). *The effect of changes in the National Assessment: Disentangling the NAEP 1985-86 reading anomaly* (Rep. No. ETS-17-TR-21). Washington, DC: National Center for Education Statistics. Retrieved August 1, 2018, from <https://files.eric.ed.gov/fulltext/ED322216.pdf>.
- Betebenner, D. W. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, 28(4), 42–51.

- Blyth, C. R. (1972). On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, 67(338), 364–366. <http://doi.org/10.2307/2284382>.
- Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items* (Vol. 4). Thousand Oaks, CA: Sage Publications.
- Ho, A. D. (2008). The Problem with "proficiency": Limitations of statistics and policy under No Child Left Behind. *Educational Researcher*, 37(6), 351–360.
- Kiplinger, V. L. (2008). Reliability of large-scale assessment and accountability systems. In K. E. Ryan & L. A. Shepard (Eds.), *The future of test based accountability* (pp. 93–114). New York: Routledge.
- Linn, R. L., & Haug, C. (2002). Stability of school-building accountability scores and gains. *Educational Evaluation and Policy Analysis*, 24(1), 29–36.
- Marion, S. F., White, C., Carlson, D., Erpenbach, W. J., Rabinowitz, S., & Sheinker, J. (2002). Making valid and reliable decisions in the determination of adequate yearly progress. In *Implementing the state accountability system requirements under the No Child Left Behind Act of 2001*. Washington, DC: Council of Chief State Schools Officers.
- NRC (National Research Council). (2010). *Getting value out of value-added: Report of a workshop*. Washington, DC: The National Academies Press.
- NRC. (2014). *Developing assessments for the Next Generation Science Standards*. Washington, DC: The National Academies Press.
- Wainer, H. (1986). Minority contributions to the SAT score turnaround: An example of Simpson's paradox. *Journal of Educational Statistics*, 11(4), 239–244.





# Comparability Within a Single Assessment System

Mark Wilson, *University of California, Berkeley*  
 Richard Wolfe, *Ontario Institute for Studies in Education, University of Toronto*<sup>1,2</sup>

## CONTENTS

INTRODUCTION .....	76
DEFINING THE CONTENT OF AN ASSESSMENT SYSTEM .....	78
Standards Defining Curriculum and Assessment Goals .....	78
Opportunity to Learn.....	79
Creating Test Specifications .....	80
Selecting Content to Implement Assessments .....	84
Targeting Assessment Content for Educational Improvement .....	85
Ensuring Coherence in Assessment Content .....	85
DESIGN OF THE CONSTRUCTS IN AN ASSESSMENT SYSTEM .....	87
Progress Variables.....	89
Designing Assessment Tasks.....	90
Categorizing Student Responses to Individual Items.....	93
Mapping Out the Construct .....	94
COMPARABILITY OF STRINGENCY OF ASSESSMENTS BETWEEN	
SUBJECT AREAS AND BETWEEN GRADES .....	100
Comparison of Tests of the Same Subject Matter Across Different Grades.....	102
Comparison of Tests on Different Subject Matters at the Same Grade Level ...	106
RELIABILITY FOR DIFFERENT USES .....	108
General Principles.....	108
Linking Error.....	108
Measurement Error.....	109
Accuracy Requirements: Total and Subscores, Individuals and Groups.....	111
General Advice .....	113

<sup>1</sup> We would like to thank Ed Haertel and Joan Herman for their review of the first draft of the chapter, as well as the rest of the committee and our fellow chapter authors for their guidance on the chapter. In particular, thank you to Ed for extensive and thoughtful edits and comments.

<sup>2</sup> The authors each contributed equally to the effort.

TRANSPARENCY IN THE DESIGN AND INTERPRETATION .....	114
For Consumers .....	114
For Evaluators .....	117
CONCLUSION .....	117
REFERENCES .....	118

## INTRODUCTION

Our interest is in comparability as it applies to large-scale assessment systems that are used for student testing and educational accountability. We focus in particular on those issues that pertain to uses within a single system, although it must be acknowledged that (a) the issues that arise are shared with a multisystem perspective, and (2) some single systems may involve multiple tests (Wilson, 1997). What questions are the assessment systems intended to answer? Such questions determine the requirements for different kinds of comparability.

We consider mainly large-scale, end-of-year testing programs, such as those provided by the Smarter Balanced Assessment Consortium (SBAC) and Partnership for Assessment of Readiness for College and Careers (PARCC) and by individual state education authorities. Large-scale assessment systems<sup>3</sup> have the principal purpose of reporting educational achievement results for all individual students and for aggregates, including classrooms, schools, districts, and whole populations (states) and subpopulations (e.g., by gender, ethnicity, etc.). The term “accountability” covers much of these purposes, although it seems a misnomer in the case of individual results and reporting: Is a student accountable for their individual achievement? Interpretation of test results at the classroom, school, and district levels needs to be tempered by consideration of student input characteristics and analyzed in light of opportunities to learn, which are factors not always included in the interpretive framework.

One of our foci is on subject-matter content and subcontent. How is the testing content domain defined and enacted in the assessment and how is it related or matched to the curriculum? How does testing content relate across curricula, such as across jurisdictions (e.g., school districts, etc.), grades, and tracks? How does it relate to classroom activity and instruction? A second focus is the comparability of the construct and the stringency of the assessments. Of course, the constructs of the assessments will be strongly delimited by the subject matter as mentioned in the previous point, but generally, the way that subject matter is defined will leave much room for further specification down to the level of the subject-matter content of the actual items in the assessments, as well as the way that the items are designed and the responses are valued and coded. Beyond this, there can also be variations in how difficult the assessments are—the “stringency” of the assessments, which we discuss further below—but generally we are referring to whether high scores and low scores are appropriately rare. How tests are centered as to difficulty and where cut scores are established are important for interpretation and use, especially between subjects and across grades, yet unless they are handled deliberately, they can be essentially arbitrary and lead to anomalous

---

<sup>3</sup> See the definition in Chapter 1 in this volume.

results, such as proficient students in one year becoming nonproficient in the next or math performance seen as great while reading is not.

The targets of assessment are primarily the students, who are the respondents and first-line recipients of the testing results (along with their parents). But scores are also aggregated and acted on at different levels: classroom, school, district, and state. And further differentiation is made according to student characteristics, such as demographics, student specialties and exceptionalities, and educational programs, with a particular concern for identifying educational coverage and gaps.

Assessment systems place critical importance on identifying and tracking trends over time. A crucial goal of an assessment system is to provide evidence about improvement in educational productivity at the state and local levels. Change is also important in the assessment of individual students, to know about achievement growth, and about how fast students are progressing toward goals (e.g., achieving proficiency).

In some assessment systems, summative end-of-year tests are complemented by “interim” or “formative” tests. If such tests and their administrative testing conditions are not seriously designed as components or estimators of the summative tests, their potential for rigorous comparability is extremely limited, and that is also true for teacher-made and school- and district-made tests. All these kinds of tests can be important tools in teaching, learning, and administration, but we do not see that issues of comparability are tractable, except where they are explicitly addressed in the test design and construction, which they rarely are. Following the calls for continuous assessment in the National Research Council’s *Knowing What Students Know* report (NRC, 2001), a number of authors did publish plans for assessment systems that might indeed deal with at least some of these design and psychometric challenges (Darling-Hammond & Pecheone, 2010; Preston & Moore, 2010; Resnick & Berger, 2010; Wilson, 2009; Wise, 2011; Zwick & Mislevy, 2011). However, little real progress has been made on the issues raised in this series of papers, and evidence seems lacking that the current interim tests from the consortia are really on the same scale as a final test or that they cumulate to a final test. And just having calibrated pools of items is nice but that does not lead to comparable scores and results without careful psychometric work. More recently, some developments are in the planning stages (Gianopoulos, 2019).

So our definition of comparability revolves mainly around the validity and accuracy of the comparisons that are intended in summative results for individuals and aggregates (classrooms, schools, districts, and states), especially across grades, subjects, and years, and in interim results where they are strongly aligned to summative tests.

In the following sections we discuss the subject-matter content, the design of the measurement constructs within the system, the stringency of the different tests within the system, the reliability of the tests with respect to different uses, and the need for transparency in the system. In each of these topics, we seek to outline what are the most important ideas and issues, to suggest ways of thinking about these issues, and to raise important questions that need to be considered regarding these issues. There are many more issues that can be raised, but we see these as particularly salient at this point in time.

## DEFINING THE CONTENT OF AN ASSESSMENT SYSTEM

What makes one test different from the others? Are the mathematics tests in SBAC, PARCC, the Trends in International Mathematics and Science Study (TIMSS), the Programme for International Student Assessment (PISA), or, for example, the Massachusetts Comprehensive Assessment interchangeable clones? Is a math test just a math test? There are obvious surface differences among these tests because of grade level, sampling, item formats, and so on. But we think there are fundamental differences in what information these and other tests can possibly provide, and a primary difference is in the definitions of their content domains, that is, of the collections of knowledge and skills (and attitudes, habits, etc.) that are accepted as the legitimate and necessary objectives of measurement and reporting in the assessment system. This may seem nonproblematic to many. But consider the following situation. A person says, "I have measured the length of this chair, and it is 21 inches long." Now, the general property is clear (linear dimension), and the unit is clear (inch), but the nature of the property is not clear—Which length is meant: the depth of the seat, the height off the ground? We see a similar problem in testing: a careful specification of the numbers and the units, but a lack of clarity of what is meant by the property under measurement. What we promote here are the following views:

1. The content of an assessment system should be given a detailed, articulated, comprehensive definition, which might be called the test content framework.
2. The origin of the framework will usually be the educational standards that are in play, but they need to be articulated, elaborated, and organized as required for defining the assessment.
3. The framework should certainly be closely related to the school curriculum frameworks, that is, to what is supposed to be taught and learned, also derived from the standards.
4. The framework should also be related to the teaching materials and methods in practice in schools.

Interpretation of assessments should pay attention to issues in alignment of these aspects. In particular, the content framework defines the basis and objective for comparability, and its connections to the standards and to curriculum and instruction provide crucial perspectives for interpreting accomplishments and gaps that comparable assessments reveal.

Without dwelling on the larger epistemology of the definition of a content domain, we can consider the practical and immediate questions of what sources are gathered in the design of an assessment to circumscribe the content domain and guide the specific content divisions and details and how these are converted or elaborated into test content frameworks.

### Standards Defining Curriculum and Assessment Goals

The basic sources for assessment content are the official state educational curricular standards. Across the United States, the Common Core State Standards Initiative (CCSSI, 2010) effectively provides curriculum goals for many assessments, although it

has now been adapted rather than adopted in many states. States or school districts can have more or less elaborated statements of their particular curricula, which incorporate the Common Core or other definitions. Before the Common Core, there were as many curricula in mathematics as there were states or maybe districts (Cogan, Schmidt, & Wiley, 2001; Schmidt, McKnight, & Raizen, 1997).

In some circumstances, such as the international TIMSS and PISA studies and early state or National Assessment of Educational Progress (NAEP) surveys, where common curricula have not coalesced—such as in social studies, arts, and perhaps science—there may be no possibility of agreeing on a single, existing curriculum, and developing and articulating a content domain has been work requiring invention, negotiation, and compromise. A special and extreme case of this is in science, where the Next Generation Science Standards (NGSS) (NRC, 2013) were developed to present a new, coherent vision of how science should be taught over grades and science disciplines. It has been recognized that materials and training for teaching science based on NGSS will take some years of development and transition, and the standards and their tests precede the instructional implementation. NGSS has been an effort to reform teaching and testing in concert, although, in the actual implementations, testing has frequently preceded the requisite extensive professional development for teachers and time needed for the implementation of this innovative curriculum (NRC, 2015).

### Opportunity to Learn

The example of NGSS implementation raises a general issue of how an assessment relates to specific teaching content and methods and to learning materials and whether students have had the opportunity to learn (OTL) the content (Wolfe, 2000). There is concern that a test cannot be fair if it deals with content for which students do not have (adequate) opportunities to learn; the interpretation of achievement differences among groups should focus in the first instance on whether there is equivalent OTL. The analysis of opportunities might be a precursor to the definition of the content framework, for example, by sampling contents from the published curricula or programs of study. Or opportunities might be measured during an assessment by asking teachers or students to indicate how instructional time was spent and which content they feel was fairly covered in instruction or, in the most traditional way, by asking students or teachers whether each item on a test had been included in instruction (Suter, 2017; Walker, 1962). We raise the suggestion that *measuring* OTL might be part of an assessment system.

However, in order to measure OTL, we need to ask the question: Where is the bridge between the documentation that defines the content and the item design and test blueprint? For example, if the source is a published curriculum, there should be a documented procedure for how the curriculum is analyzed, transformed, and sampled to arrive at the test specifications. A comprehensive system for this is given by the Study of Mathematics and Science Opportunities associated with TIMSS-1995 (see Schmidt et al., 2001). A unified classification and coding system was defined for the content of curricula, textbooks, and test items and tests. It is a three-dimensional system. Dimension 1 is the content, a hierarchical list of topics and subtopics; dimension 2 is performance expectations, a hierarchical list with main categories of knowing, using routine procedures, investigation and problem solving, mathematical reasoning, and



communicating; and dimension 3 is perspectives, including attitudes, careers, participation by underrepresented groups, and habits of mind. The coding is designed to be multidimensional but it is also multiattributational because a given unit (textbook block, test item, or curriculum specification) is coded on all three dimensions and potentially on multiple aspects within each dimension. This is illustrated in Figure 4-1.

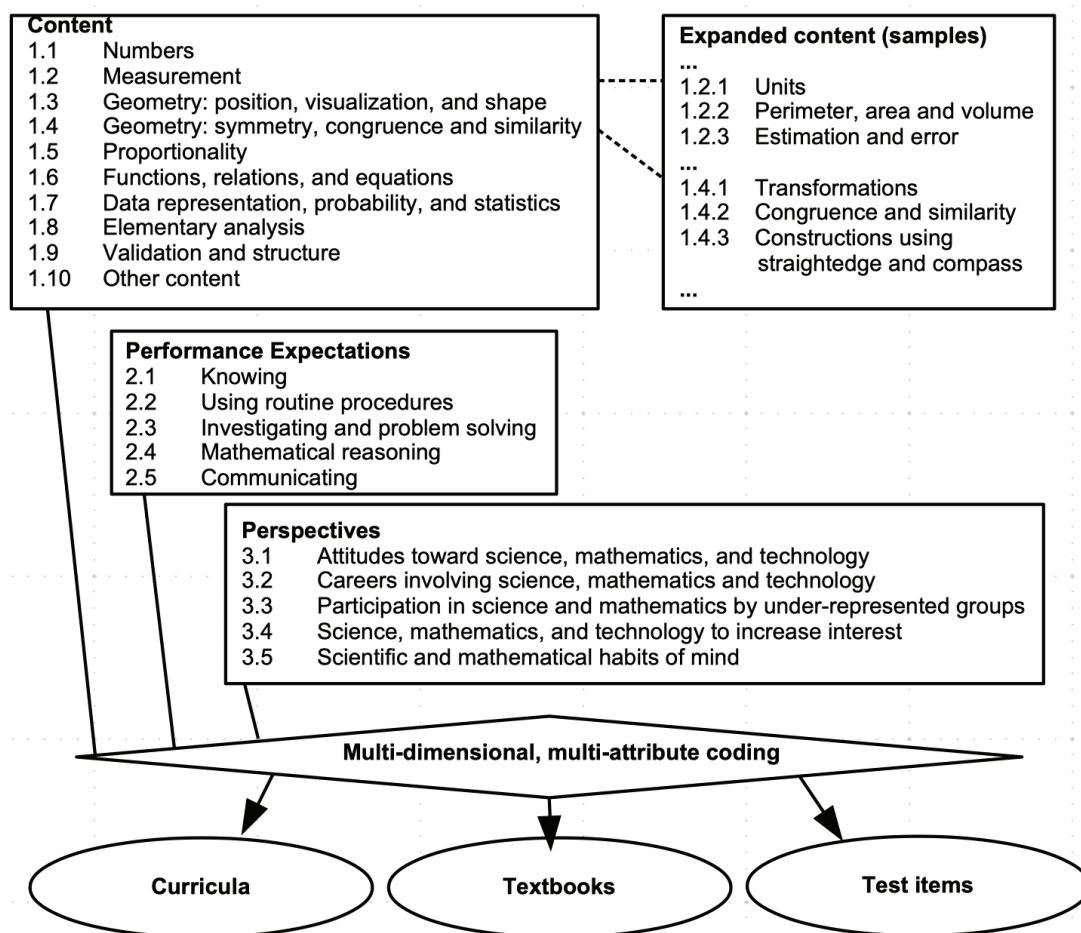


FIGURE 4-1 TIMSS 1995 mathematics framework for coding curricula, textbooks, and tests.

### Creating Test Specifications

Assessment *blueprints* are supposed to be the formal instructions for item writers and test form constructors to create a test. They provide written documentation defining in detail the content, content breakdowns, and assessment goals (e.g., item types). Are these blueprints sufficiently explicit and detailed to allow replication of comparable tests?

In assessment systems based on Common Core, the beginning points are lists of *standards*, which are statements for a particular grade and area of content of what

students should be able to do or demonstrate.<sup>4</sup> The standards can be divided into final standards, what students should ideally be able to do at the end of a successful course of study, and enabling standards, which are intermediate achievements and steps toward the final standards. A conceptual problem with standards is that they often seem to be binary—met or not met. It is difficult to see how to use them in describing partial or graded accomplishments. A student meets a standard or not. How can one partly meet a standard? Presumably the answer is that the content surrounding or composing a standard is actually much finer grained, including subcontents, depths of knowledge, and performance expectations. We can ask *how well* a student meets a standard only by employing a variety and gradation of test items. This creates an important technical challenge for test developers, because the available test space (number of items in play in an assessment) is limited and relatively small considering (1) the number of subcontents and standards that are to be included and (2) the item density required to measure them comprehensively.

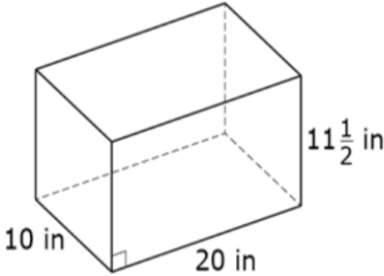
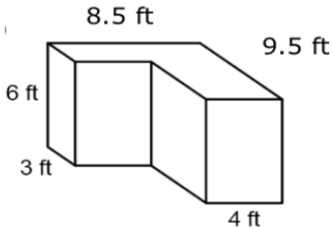
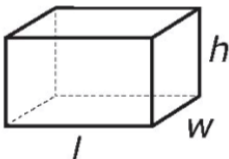
In Figure 4-2a is an example of a grade 6 geometry standard from the Common Core together with three released items from the SBAC pool that are coded, by SBAC, to different levels of depth of knowledge. The standard is labeled in the Common Core as having students “solve real-world and mathematical problems involving area, surface area, and volume.” Its statement specifies the kind of mathematical reasoning students are supposed to have, namely, relating packing unit cubes with the formula for volume. It is not obvious how these example items represent “real-world” problems, so they must be “mathematical” problems, and there is nothing in the items that would require or demonstrate understanding of the underlying unit-cube-packing basis of the formula. The first item is a straightforward application of the formula. The second is complicated by requiring some geometric insight into decomposing the given shape into two prisms and deducing their dimensions. The third involves a logical trick, or a sequence of deductions, that has little to do with the content standard. How many additional items would it take to fully measure student achievement in all aspects of this standard, for an individual student or for an aggregate group of students?

Another Smarter Balanced grade 6 mathematics item, given in Figure 4-2b, provides a more complete measurement of this standard. In this case, the context is “real world.” The student is to figure the volume of a rectangular prism—the cargo hold in a truck—by seeing how cubes of a particular size would pack the width, height, and depth of the hold. The formula  $V = lwh$  is not explicitly invoked or required but probably occurs naturally, as in the sample response. Because the student is asked to show how the answer was determined, there can be evidence about the understanding of the geometry. The scoring rubric for this extended response item awards points for a correct explanation as well as for the correct answer.

The relatively small set of standards in Common Core is best thought of as defining major mileposts in achievement in the content domain. As we look over a collection of contents and standards, we think about how well a student does in the domain, on the average or over standards. The number of standards in play and the need for multiple items per standard creates an issue for test design; for example, the Common Core

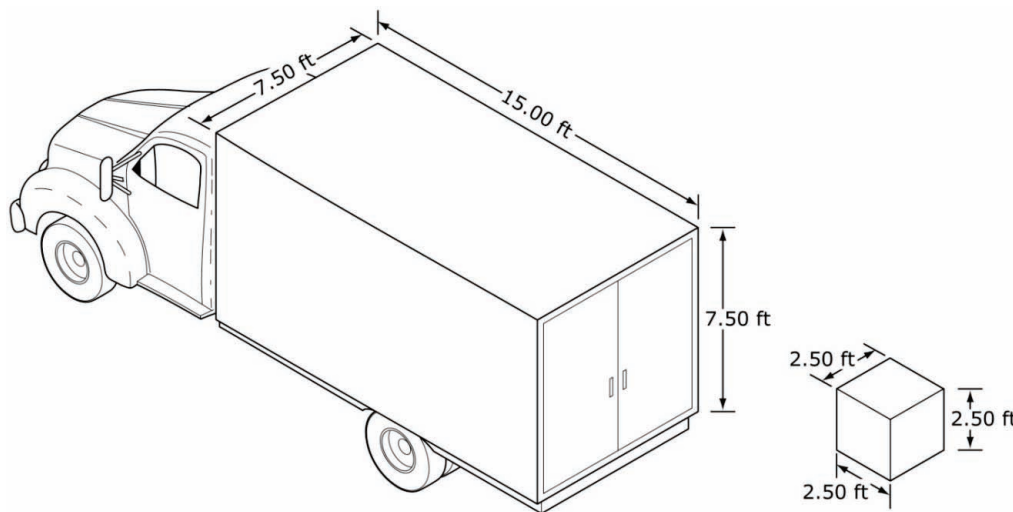
---

<sup>4</sup> An earlier educational policy view of this was the definition of *behavioral objectives* (Ebel, 1970; Tyler, 1934).

<p>CCSS MATH CONTENT STANDARD 6.G.A.2</p>	<p>Find the volume of a right rectangular prism with fractional edge lengths by packing it with unit cubes of the appropriate unit fraction edge lengths, and show that the volume is the same as would be found by multiplying the edge lengths of the prism. Apply the formulas <math>V = l w h</math> and <math>V = b h</math> to find volumes of right rectangular prisms with fractional edge lengths in the context of solving real-world and mathematical problem</p>	
	<p>Consider this figure.</p> <p>Enter the volume, in cubic inches, of the right rectangular prism.</p>	<p>DOK Level 1</p>
	<p>This figure was created by joining two right rectangular prisms.</p> <p>Enter the volume, in cubic feet, of the figure.</p>	<p>DOK Level 2</p>
 <p>(Not drawn to scale)</p>	<p>A right rectangular prism has a height of 5 centimeters. Is it possible that the volume of the prism is 42 cubic centimeters?</p> <p><u>If it is possible:</u></p> <p>Enter a possible length and width, in cm, of a prism with a height of 5 cm.</p> <p><u>If it is <b>not</b> possible:</u></p> <p>Enter a possible volume (in cubic centimeters) and the corresponding length and width (in centimeters).</p>	<p>DOK Level 3</p>

**FIGURE 4-2a** A Common Core standard in grade 6 geometry and three released items from SBAC at different levels of depth of knowledge (DOK).

Cube-shaped boxes will be loaded into the cargo hold of a truck. The cargo hold of the truck is in the shape of a rectangular prism. The edges of each box measure 2.50 feet and the dimensions of the cargo hold are 7.50 feet by 15.00 feet by 7.50 feet, as shown below.



What is the volume, in cubic feet, of each box?

Determine the number of boxes that will completely fill the cargo hold of the truck. Use words and/or numbers to show how you determined your answer.

*Sample Top-Score Response:*

The volume of each box is 15.625 cubic feet.

54 boxes completely fill the cargo hold of the truck. The length of the cargo hold is 15 feet, so 15 divided by 2.50 equals 6. The width and height of the cargo hold are each 7.5 feet, so 7.5 divided by 2.5 equals 3. So the 6 boxes times 3 boxes times 3 boxes equals 54 total boxes that fit in the cargo hold.

**FIGURE 4-2b** A released constructed-response item from SBAC for the same Common Core standard in grade 6 geometry.

seems to have about 25 standards per grade area per subject area. Accurate, reportable measurement of achievement for one standard might require, say, seven multiple-choice items, or perhaps fewer constructed-response items. That translates to a very long test if our goal is really to measure each standard, one by one, so clearly some kind of sampling, rotation, and averaging will be necessary to “cover” the content in an appropriate way. We return to this matter in a later discussion of reliability.

Regardless of the source of the content definition, eventually there has to be some list or table of contents for the purposes of defining the blueprint and specifying how many items of what kind will be used in the test. This might be a single list of contents, but more likely it will be a hierarchical list of contents, where the main contents are divided into subcontents. Those will potentially define subtests and subscores. Traditionally, the content list, possibly hierarchically arranged, is considered one dimension of a matrix where contents are the rows and the columns are often cognitive levels (e.g., derived from Bloom’s taxonomy) (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956). Other column dimensions are possible, such as types of skills within the content area, or item type (multiple choice, constructed response, and performance). The columns can be multifactor, such as cognitive level  $\times$  item type. Consider the specification for the NGSS where three dimensions are crossed: science and engineering practices, crosscutting concepts, and disciplinary core ideas. Each performance expectation (i.e., content or standard) is categorized on the three dimensions.

In summary, a conventional test blueprint is supposed to provide an organized and detailed specification of the contents to be measured, usually represented as a matrix of content category by other types of categories. Each cell corresponds to one point in the defined content domain. It has an associated weight or *quota* indicating the number of items, or the number of score points in the test to be obtained for the cell. These numbers may be fixed or presented in some kind of design for sampling or rotation (e.g., over forms or over time). The quotas must be adequate for the total test and for any subscores that are to be reported.<sup>5</sup>

### Selecting Content to Implement Assessments

Is the content of an assessment system, in the end-of-year comprehensive test or in the combination of end-of-year and interim tests, intended to be comprehensive or selective with respect to the content domain? That is, is the intention to cover (in theory, with sampling, and over time) the whole domain, or is there a priori selection or filtering of content? Some filtering might come from omission; that is, certain contents are simply not in the blueprint matrix (hidden omission). Other filtering might come from zeros appearing as the blueprint quota (explicit omission, although only for those who know the blueprint quota). In either case, the content is forgone and lost to the assessment system. For example, in the past, writing tests have sometimes involved no writing by students; mathematics tests have lacked extended problem solving; and science topics or concepts may be omitted. Obviously, if important content is omitted, it is not part of the content domain being tested, and some adjustment or restatement

---

<sup>5</sup>Sampling accuracy for total scores and subscores depends on the number of items by content category for all types of test design—single form, matrix-sampled forms, multistage tests, and computer-adaptive tests—although the effects vary between individual and aggregate scores and between the different test designs.

must be made of the name and description of the domain. In a test in which there is mathematics without problem solving, science without laboratory work, and writing mechanics without composition, using the simple traditional labels would seem to be misleading.

If some standards are not measured in sufficient detail or not measured at all, can we claim that the assessment evaluates achievement of the standards? If content selection varies over time or over grade, then claims of comparability are suspect. For example, if actual writing samples or extended mathematics investigations are included only in selected grades, then vertical comparability in writing and mathematics over grades is not obtained. If science topics for a grade are rotated across years, then the content definition of science achievement varies and horizontal (year-to-year) comparability may be lost.<sup>6</sup>

### Targeting Assessment Content for Educational Improvement

Is the underlying philosophy of content of an assessment system intended to follow or lead educational change? That is, an assessment may be intended to closely mirror current standards and current curriculum and to reflect current classroom practices, or it might reflect aspirations for how teaching and learning in a subject should be developing and moving. Considering the problem of assessing science in the current moment of transition to the NGSS curriculum, in most schools, implementation of NGSS is partial at best. So should we define science assessment content according to NGSS, which would lead to cries of student lack of OTL, or to current content and practices, which are regarded as outmoded? This can create a dilemma of comparability because change over time or differences over jurisdictions are tied to discrepancy between constant test content and variable curricular content.

### Ensuring Coherence in Assessment Content

An assessment system is *coherent* (NRC, 2001, pp. 255–256) if it is based on a logical and consistent definition of the content domain and there is a rigorous connection between the domain and the technical design of the assessment, including the tests, analyses, and reporting, with particular reference to models for instruction and student learning. Without coherence it would be difficult to prepare or to evaluate comparable measurements between tests or over time.

A coherent content definition is needed to justify naming the assessment results according to the domains and subdomains, which should be appropriately qualified according to the selections and filters applied in the test specification and to the role of the targets in setting the specifications (as described above). If this is done honestly, the limitations of

---

<sup>6</sup>It may be argued that selection or rotation of content does not always limit the possibility for fundamental comparability if we can suppose that the specific contents are only examples of essential knowledge and skills that will be demonstrated in different contexts and contents. For example, essential mathematical reasoning, we might think, will be shown and measured in different mathematical topic areas (algebra, geometry, etc.), or reading comprehension ability can be effectively observed with different types of reading materials (informational, narrative, etc.). But why would the content design include a broad, articulated definition of content if a more narrow one would work as well? It also ignores the question of whether items designed for the subcontents will vary systematically in terms of their measurement effects and, hence, give systematically different results depending on which areas are rotated in and out.



an assessment will be evident. Consider, for example, the descriptions “mathematics in grade 8 according to the Common Core including basic knowledge, skill, and applications but excluding extended/practical problem solving”; “science for middle school according to NGSS content domains excluding biology”; and “literal comprehension of grade-level materials without inference and connection to real-world life.” The larger domains are traditionally named mathematics, science, and language arts or reading, but those shorter titles can be overblown and misleading in these circumstances.

In most assessments, there are subdomains in the content definition and correspondingly subscores are expected in the reporting, for example, arithmetic, geometry, and algebra within “mathematics”; word recognition, inference, and main idea within “reading”; and biology, physics, and chemistry within “science.” These kinds of divisions are used in the definition of the overall content, and they are part of the stratification of the test specification and item selection. However, given the presence of these subdomains in the materials associated with the test, it is inevitable that administrators and teachers will want to see the scores on these subdomains. Unfortunately, there is no guarantee that the test will produce accurate subscores for individuals or for aggregates (classrooms, school, districts, states, ethnic groups, etc.) unless that is an explicit goal of the assessment design and considerable investment is made in sampling content, that is, in number of items per subarea. The same logic applies to specific or general “claims” (i.e., aside from subdomains) to be measured. Is the list of claims intended to be exemplary or comprehensive?

If the assessment goals do in fact include a specification for scores and reports by subcontents or claims, then each corresponding subtest needs to be given a full measurement and statistical treatment to ensure that it has content coherence. The question also arises of how the subdomain and claim results are connected to the overall results. For example, in NAEP mathematics assessments, the overall results have been defined as the (weighted) sums of the results of five subdomains, each of which has been measured with high accuracy and reported separately (NCES, 2018). This high standard of test development is seldom reached in other testing programs.

If the assessment goals include producing subtest scores and reports, this must be evaluated from the perspective of comparability. In particular, are the subscore results at the individual level comparable to the parallel results obtained at the aggregate level? Are the subscore results from one assessment year comparable to those in the next year?<sup>7</sup>

---

<sup>7</sup> These points seem obvious on their face, but in many assessment systems the subscores reported for individuals or groups are merely *relative* indicators—the difference between performance on items in the subarea and performance overall. They are sometimes called “relative strengths and weakness.” By definition, over the population the balance must be zero; this is a zero-sum game. They cannot be considered serious measures of the subdomains. There are no population results and nothing to compare over time, as admitted from the California reporting system for SBAC:

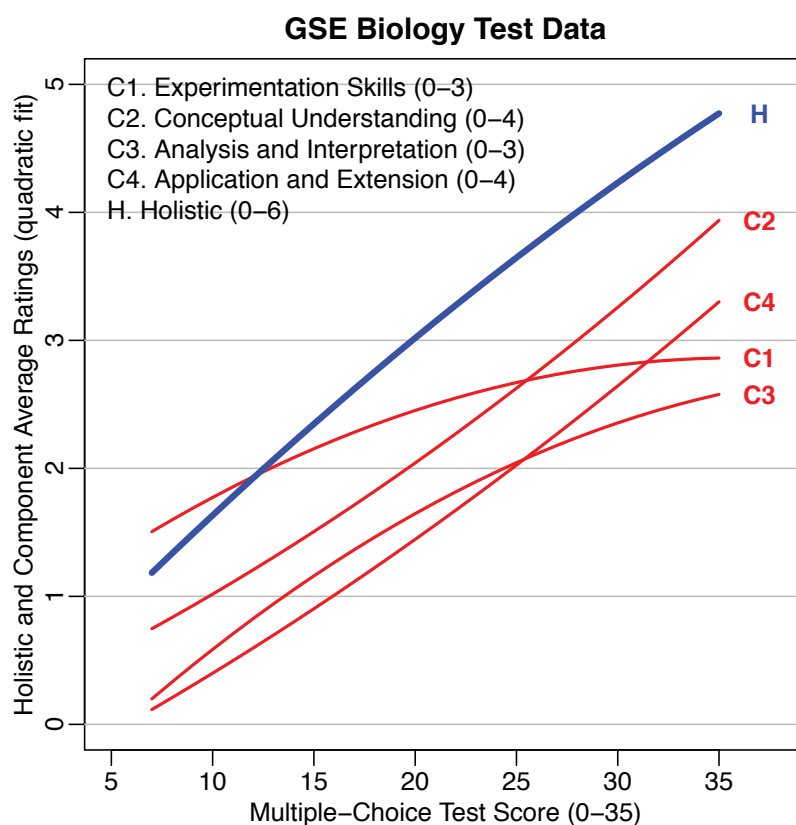
Unlike the overall ELA or mathematics score, the Assessment Target Report does not address absolute performance, but rather the report is an indicator of strengths and weaknesses relative to the test performances as a whole of the group you are viewing. Unlike performance levels provided at the content area level, these strengths and weaknesses do not imply proficiency or that a particular content standard has been met. The target report shows how a group of students performed on a target compared to their overall performance on the assessment. (CAASPP, 2015)

An implicitly unidimensional design and analysis formulation is rarely going to be anything other than an approximation to the complexity of the content domain and the measurements made in a test. Consider, for example, the case of the California Golden State Exam (GSE) biology test (this discontinued testing program is described further in the next section; see Bennett & Carlson, 1986), there were 35 multiple-choice items covering knowledge across the biology curriculum. Also, students did a practical laboratory experiment. In one year, this concerned “rice blight” and involved examining disease resistance in rice plants, determining genetic relationships in rice plants, and testing for flavor compounds. The students had to show experimental skill with laboratory equipment, write out their conceptual understanding of what they found about disease resistance, analyze and interpret information from genetic tabulations of resistance and of flavor in rice populations and generations, and finally write an essay about the implications of the findings for the future of land where disease-resistant rice is grown but then abandoned for 25 years. The student report on the experiment was rated by science teachers for four components, on rating scales that varied from 0–3 to 0–4 and on a wholistic rating scale of 0–6. As part of a special study, a wholistic rating was designed as a general progress variable concerning what one hopes to see as students learn biological science. The four component scales were also ratings of progress a student would be expected to make on different aspects of biology: experimental skill, comprehension and understanding, analysis and interpreting, and application and extension.

Altogether, the testing record for scoring and reporting included the score from the multiple-choice test, the four component scores, and the wholistic score. This could in theory have been reported as five distinct results. The results are correlated as shown in Figure 4-3, where the X axis corresponds to the multiple-choice score (knowledge) and the Y axis corresponds to the component and wholistic ratings of the student’s experimental work. It is interesting to see that the wholistic rating and two of the components, understanding and application, seem to increase linearly with the multiple-choice knowledge score, while the other two components, experimental skill and analysis and interpretation, show a diminishing relationship to the knowledge embodied in the multiple-choice items. This may be a ceiling effect for these two components. But it is reasonable to conclude that there are subdomains corresponding to the four component scores and these are correlated with but not identical to the general domain measured by the wholistic score and the multiple-choice score. But this is not to say that we could obtain accurate, replicable subscores.

## DESIGN OF THE CONSTRUCTS IN AN ASSESSMENT SYSTEM

The focus of the previous section was on the *domain* perspective on test content. From this perspective, the content of a test is defined by listing the specific parts of the subject matter that are to be included in the test. These may be derived from curriculum documents (sometimes called curriculum “frameworks”) or from textbooks or other content-defining sources. Sometimes these will appear in the form of a list of topics, or sometimes as a table, usually called a “test blueprint” showing how these topics can be distributed across other categories such as grade level, depth of knowledge, or a set of skills that have been hypothesized to span the content that is being assessed.



**FIGURE 4-3** Relationship of scores from the wholistic rating, component ratings, and multiple-choice scores in a GSE biology experiment.

This section focuses on an alternative perspective for defining test content: the *construct* perspective. Under this approach, learning is conceptualized not simply as a matter of acquiring quantitatively *more* knowledge and skills, but as progress toward higher levels of competence as new knowledge is linked to existing knowledge and as deeper understandings are developed from and take the place of earlier understandings. Certainly, this perspective is related to theoretical views of the domains (biological science, linguistics, sociology, etc.), but it is derived specifically from research into the underlying cognitive structure of the domain—how knowledge, understanding, and skill in the domain are formed in the minds of students—and from research-based analysis and conclusions about what constitutes higher and lower levels of performance or competence. In particular, this perspective can help in understanding the importance and power of cross-grade comparisons, vertical scaling, and related claims about score scales, and it can also help in developing successful examples of these. This perspective derived in part from research into the underlying cognitive structure of the domain and in part from the judgments of professional educators about what constitutes higher and lower levels of performance or competence. But it should also be informed by empirical research into how students respond to instruction or perform in practice (NRC, 2001, pp. 218–219). The empirically grounded and criterion-referenced interpretation of

student performance is a central basis for—or the definition of—comparability within an assessment system.

### Progress Variables

There are several ways that such constructs can be conceptualized, using structures such as dimensions, classes, and networks. In this section, we concentrate on the possibilities for dimensional structures, including multidimensional as well as unidimensional structures. We refer to these dimensions as progress variables because they embody a developmental perspective on the assessment of student achievement and growth. The term *variable* is derived from the measurement concept of focusing on one important characteristic to be measured at a time. A progress variable is a well-thought-out and researched hierarchy of qualitatively different levels of performance. Thus, a progress variable defines what is to be measured or assessed in terms general enough to be interpretable across a curriculum or state testing program but specific enough to guide the formation of the other components. When instructional objectives are linked to the variable, then it also helps define what is to be taught. As noted above, there will usually be multiple progress variables in a single curriculum. Progress variables are one model of how assessments can be integrated with instruction and accountability. Progress variables provide a way for large-scale assessments to be linked in a principled way to what students are learning in classrooms while remaining independent of the content of a specific curriculum. Thus, they are a potential solution to the problem mentioned above (in the section “Defining the Content of an Assessment System”) that many educational systems are composed of amorphous and variable curricula, whereas assessments are desired (and most often required) to have uniformity, not least for the purpose of fairness.

This approach assumes that, within a given curriculum, student performance on curricular variables can be traced over the course of the year and across grade levels, facilitating a more developmental perspective on student learning. Assessing the growth of students’ understanding of particular concepts and skills requires a model of how student learning develops over a set period of (instructional) time. A growth perspective helps one move away from “one-shot” testing situations and away from cross-sectional approaches to defining student performance, toward an approach that focuses on the process of learning and on an individual’s progress through that process. Clear definitions of what students are expected to learn, and a theoretical framework of how that learning is expected to unfold as the student progresses through the instructional material, are necessary to establish a developmental perspective on the construct validity of an assessment system (NRC, 2001, p. 292).

Rather than focusing on an item-by-item or standard-by-standard content match, progress variables allow the matching of sets of tasks to overarching frameworks. For example, if a progress variable such as “designing and conducting investigations” is well represented in a state- or district-level assessment, one can have confidence that content standards relating to science as inquiry are being measured by that assessment.

It is important to distinguish the activity described above from an ancillary practice associated with most assessment tests: the setting of reporting categories called “performance levels” or assessment level descriptions. The terminology conventionally used

in naming performance levels can be overly general, which tends to mystify the reader rather than make concrete reference to what a student can do. For example, in PARCC (n.d.) the performance levels are described as follows:

- Level 1: Did not yet meet expectations.
- Level 2: Partially met expectations.
- Level 3: Approached expectations.
- Level 4: Met expectations.
- Level 5: Exceeded expectations.

The semantics of this list are not obvious. What is the difference between “Partially met” and “Approached”? Does “Did not yet meet” mean that there was zero achievement, less than partial? How does one “Exceed” expectations? That would seem to indicate higher expectations. Remember that these same level descriptors are used for many different contents and subcontents, but, in fact, they possibly make sense only if the terms “not yet met,” “partially met,” “approached,” and “exceeded” are explained in terms of the specific content at hand. There is often a process of performance-level description as another step in content definition, which will add verbal descriptions of the levels, although often these too are also vague and lacking in concrete educational referents (Glass, 1978; Shepard, 1980).

### *An Example of a System Based on Progress Variables: The Golden State Exams*

In order to avoid either unduly praising a specific assessment system or unduly blaming such, we use as an example throughout this chapter a now-discontinued set of assessments called the Golden State Exams (Bennett & Carlson, 1986). The GSE program in the state of California consisted of a set of high school honors examinations. These were end-of-course examinations in a number of subjects, including mathematics (Algebra, Geometry, and High School Mathematics), language (Reading & Literature, Written Composition, and Spanish Language), science (Physics, Chemistry, Biology, and Coordinated Science), and the social sciences (U.S. History, Government & Civics, and Economics). Each examination consisted of a set of multiple-choice items and at least one written response item.

Based on their scores on a particular GSE, examinees were categorized into one of six hierarchically ordered performance levels—descriptive categories of student performance in each subject area. Figure 4-4 contains these categories for Algebra as an example. The top three levels (4, 5, and 6) were considered “honors” levels (School Recognition, Honors, and High Honors, respectively).

### **Designing Assessment Tasks**

Assessment tasks create the match between classroom instruction and the constructs underlying the assessments. The critical element to ensure is that each question (or item) in the assessment task is matched to at least one construct; more explicitly, responses to the question can be mapped to specific levels of the construct map. This coherence is engendered by adherence of the task design to a construct map. This is what gives tasks the developmental coherence that allows for them to be different in their design

<p>Level 6. Student work demonstrates evidence of rigorous and in-depth understanding of mathematical ideas; the work:</p> <ul style="list-style-type: none"> <li>• Is consistently correct and complete, and shows thorough understanding of mathematical content and concepts</li> <li>• Communicates clear and logical explanations of solutions to problems that are fully supported by mathematical evidence</li> <li>• Shows problem-solving skills that include appropriate generalizations, connections, and extensions of mathematical concepts</li> <li>• Includes effective use of mathematical language, diagrams, graphs, and/or pictures</li> <li>• Shows skillful and accurate use of mathematical tools and procedures, often with multiple and/or unique approaches</li> </ul>
<p>Level 5. Student work demonstrates evidence of solid and full understanding of mathematical ideas; the work:</p> <ul style="list-style-type: none"> <li>• Is essentially correct and complete, although it may contain minor flaws</li> <li>• Communicates explanations of solutions that are supported by mathematical evidence</li> <li>• Shows problem-solving skills that include connections and extensions of mathematical concepts</li> <li>• Shows appropriate use of mathematical language, diagrams, graphs, and/or pictures</li> <li>• Includes accurate use of mathematical tools and procedures</li> </ul>
<p>Level 4. Student work demonstrates evidence of substantial understanding of mathematical ideas; the work:</p> <ul style="list-style-type: none"> <li>• Is usually correct and complete, although it may contain flaws</li> <li>• Communicates explanations of solutions that are supported by mathematical evidence for most tasks</li> <li>• May contain evidence of problem solving without connecting or extending mathematical concepts</li> <li>• Includes frequent use of mathematical language, diagrams, graphs, and/or pictures</li> <li>• Usually shows evidence of appropriate use of mathematical tools and procedures</li> </ul>
<p>Level 3. Student work demonstrates evidence of a basic understanding of mathematical ideas; the work:</p> <ul style="list-style-type: none"> <li>• Is sometimes correct; however, it may lack either depth across the mathematical content areas or may show gaps in understanding of some concepts</li> <li>• Communicates explanations of solutions that are supported by mathematical evidence for some tasks, but explanations are very weak or missing for other tasks</li> <li>• May show ineffective or inconsistent problem solving</li> <li>• Shows some evidence of use of mathematical language, diagrams, graphs, and/or pictures</li> <li>• Shows some appropriate use of mathematical tools and/or procedures for some tasks</li> </ul>
<p>Level 2. Student work demonstrates evidence of limited understanding of mathematical ideas; the work:</p> <ul style="list-style-type: none"> <li>• Shows little evidence of correct solutions and is incomplete</li> <li>• Provides limited explanations of solutions that are not supported by mathematical evidence</li> <li>• Shows limited evidence of problem-solving, arithmetic computations may be correct but unrelated to the problem</li> <li>• Shows limited evidence of use of appropriate mathematical language, diagrams, graphs, and/or pictures</li> <li>• Includes limited or inappropriate use of mathematical tools and procedures</li> </ul>
<p>Level 1. Student work demonstrates little or no evidence of understanding of mathematical ideas; the work:</p> <ul style="list-style-type: none"> <li>• Is rarely correct and has major mathematical errors</li> <li>• Provides little or no explanations of solutions</li> <li>• Shows little or no evidence of problem solving</li> <li>• Shows little or no evidence of the use of appropriate mathematical language, diagrams, graphs, and/or pictures</li> <li>• Includes little correct or appropriate use of mathematical tools and /or procedures</li> </ul>

FIGURE 4-4 GSE performance-level descriptions for algebra.



and hence have specialized uses within the system, yet still be a coherent part of that system.

Explicitly aligning the instruction and assessment addresses the issue of the content validity of the assessment system (see the section “Defining the Content of an Assessment System”), as well. Traditional testing practices—in standardized tests as well as in teacher-made tests—have long been criticized for oversampling items that assess only basic levels of knowledge of content and ignoring more complex levels of understanding (e.g., Schmidt et al., 1997). Relying on progress variables to determine what skills are to be assessed means that assessments focus on what is important, not what is easy to assess. Again, this reinforces the central instructional objectives of a course. Resnick and Resnick (1992) argued: “Assessments must be designed so that when teachers do the natural thing—that is, prepare their students to perform well—they will exercise the kinds of abilities and develop the kinds of skill and knowledge that are the real goals of educational reform” (p. 59). Variables that embody the aims of instruction (e.g., “standards”) can guide assessment to do just what Resnick and Resnick were demanding. In a large-scale assessment, the notion of a progress variable is more useful to the parties involved than simple number-correct scores or standings relative to some norming population, enriching the possibilities for planning and interpretation.

A variety of different task types may be used in this assessment system, based on the requirements of the particular situation. They can be of different designs, reacting to specific educational needs for the assessments, but, because they still relate to the underlying construct (and its levels), they can be used within the system in a coherent way. Note that this developmental coherence does not guarantee psychometric uniformity—that issue must still be addressed empirically. There has always been a tension in assessment design between the use of multiple-choice items, which are perceived to contribute to a more reliable assessment, and other, alternative forms of assessment, which are perceived to contribute to the “authenticity” of the assessment and, hence, to the validity of test interpretation. Specifically, one can distinguish two major characteristics of the assessment design space: (1) control over task specification (extremes being externally prescribed tasks versus the “ad hoc” tasks that a teacher develops to meet the needs of students) and (2) control over judgment (extremes being machine scorable versus a teacher giving an overall rating or “grade” based on their judgment). The point is not that testing situations with high or low levels of control are better, but that various tasks with varying levels of control must be designed to meet the varying assessment needs of classrooms, schools, and districts, and may be best deployed as a mixture of types.

In large-scale testing situations, the basis on which the mix of task types is decided may be somewhat different from that in embedded assessment contexts. Again, this need not challenge the developmental coherence of the system, so long as the tasks at both levels are consistent with the constructs and their levels. Many large-scale tests are subject to tight constraints both in terms of the time available for testing and in terms of the financial resources available for scoring. Thus, although performance assessments are valued because of their perceived high validity, it will likely not be possible to collect enough information through performance assessments alone to accurately estimate each examinee’s proficiency level; multiple-choice items, which require less

time to answer and which may be scored by machine rather than by human raters, may be used to increase the reliability of the large-scale test.

For example, the GSEs each contained both a set of multiple-choice items and at least one open-ended item. The multiple-choice items were each designed to assess a specific performance (construct) level (see Figure 4-4). In addition, most of the GSEs contained two performance assessments—either extended written-response items or, for science for example, an extended series of responses to a scientific exploration. The crucial link is that these scores in the scoring guides were also matched to the performance levels. This gives the test, even though it is composed of two different types of tasks, coherence in its relation to the underlying construct.

### **Categorizing Student Responses to Individual Items**

The outcome space is the set of ordered categorical outcomes into which student responses are to be categorized for each of the individual items associated with the levels of a particular progress variable. This applies in different ways for different types of items: in multiple-choice items, the students self-select their responses, which are linked to a construct level, while in written and other complex responses, student responses are judged by expert raters (and may also be machine scored based on those expert judgments). This is common across many other types of assessments too—with computer adaptive test (CAT) administrations, diagnostic classification models, or tests based on the construct perspective more broadly. In practice, these are presented as scoring guides for student responses to assessment tasks. These are supplemented by *exemplars*: examples of student work at every scoring level for every task and variable combination.

For the information from assessment opportunities to be useful to teachers, it must be couched in terms that are directly interpretable with respect to the instructional goals of the variables. Moreover, this must be done in a way that is intellectually sound and practically efficient. Scoring guides have been designed to meet these two criteria. A scoring guide serves as a practical definition for a variable by describing the performance criteria necessary to achieve each score level of the variable. For an example of a method of designing scoring guides consistent with this approach, see Wilson (2005, Chapter 4).

The scoring guides are meant to help make the performance criteria for the assessments clear and explicit (or “transparent and open” to use Glaser’s [1990] terms), not only to the teachers but also to the students, parents, administrators, and/or other consumers of assessment results. In fact, we strongly recommend to teachers that they share the scoring guides with the students as a way of teaching students what types of cognitive performance are expected and to model the desired processes. Although a little uncomfortable with this at first, because it could be construed as “teaching to the test” or “giving students the answers,” many teachers found that explicit discussions of what they expected and of how students could improve their performance could be a useful pedagogical tool. In some classrooms, teachers have taught students to score their own (or their partners’) work using modified scoring guides. Students appreciate this sharing of the assessment approach:

She [the teacher] gave us a chance to see what we did right and what we did wrong. You really can understand the work you're doing. (Roberts & Sipusic, 1999)

They also found out what other students were thinking:

You learn how different students can have different scores even though they're from the same classroom and have the same teacher. You can see what their understanding and knowledge is and you can compare it to your own understanding and knowledge. (Roberts & Sipusic, 1999)

Because there will inevitably be questions of interpretation when applying a scoring guide to a particular task, especially for teachers who are new to using the assessment system, it is recommended to supplement the scoring guides with exemplars. These are actual samples of student work, selected by teachers, to illustrate typical responses for each score level for specific assessment activities, and accompanied by brief explanations of what to note.

The idea of scoring guides is not new in large-scale testing; however, "rubrics" are often written to be item specific rather than being based on a more general underlying structure. In addition, a form of exemplar (referred to in the GSE program as an "anchor paper") is quite often provided for the raters of written-response items in large-scale testing contexts.

The existence of scoring guides can be an advantage even when there is no explicit need for them. Multiple-choice items do not need a scoring guide for scoring, but indeed something very like a scoring guide is important when developing multiple-choice items, for both the question itself and the distractors. Of course, development of a scoring guide should be an essential step in developing open-ended prompts.

### Mapping Out the Construct

A *Wright map* (Wright, 1977) is a graphical and empirical representation of a progress variable, showing how it unfolds or evolves over time in terms of student performance (see Figure 4-5 for an example). Wright maps are just one of several similar approaches. For example, NAEP's "Reckase charts" and the earlier "scale anchoring" are well-known alternatives (Beaton & Allen, 1992). A Wright map is derived from empirical analyses of student data on sets of assessment tasks. It is based on an expected ordering of these assessment tasks from relatively easy tasks to more difficult and complex ones. A key feature of such a map is that both students and tasks can be located on the same scale, giving student proficiency the possibility of substantive interpretation, in terms of what the student knows and can do, and where the student is having difficulty. This substantive and criterion-referenced interpretation based on the construct map is the bedrock for comparability in using the results from the tests.

A Wright map embodies two advantages over the traditional method of reporting student performance as total scores or percentages. For one, it allows teachers to interpret a student's proficiency in terms of average or typical performance on representative assessment activities; second, it takes into consideration the relative difficulties of the tasks involved in assessing student proficiency.

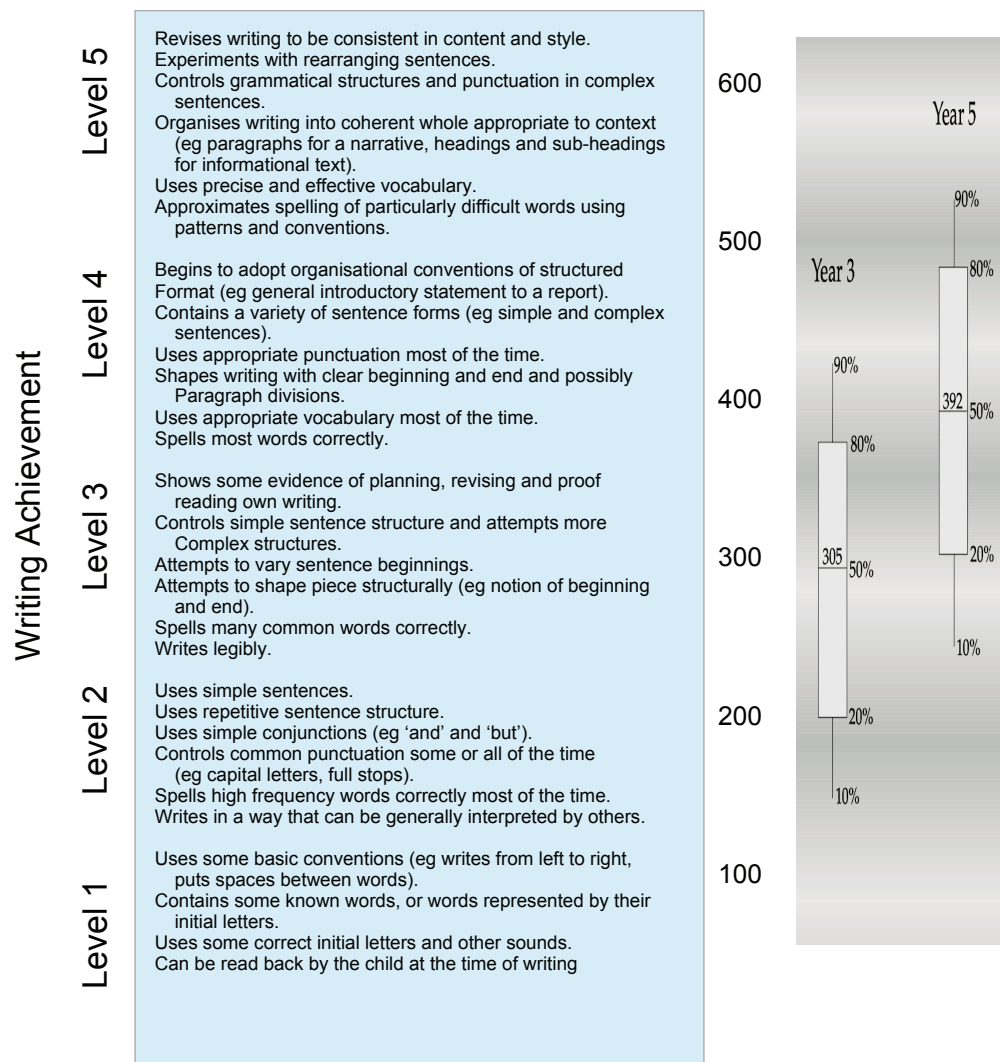


FIGURE 4-5 Wright maps of Australian National Literacy Survey.

Once constructed, a Wright map can be used to record and track student progress and to illustrate the skills a student has mastered and those that the student is working on. By placing students' performance on the continuum defined by the map, teachers can demonstrate student progress with respect to the standards that are inherent in the progress variables. Such maps, therefore, are one tool to provide feedback on how the class as a whole is progressing. They are also a source of information to use in providing feedback to individual students on their own performance.

Wright maps are available in many forms, and have many uses in classroom and other educational contexts, as well as beyond the classroom. The maps can be very useful in large-scale assessments, providing information that is not readily available through numerical score averages and other traditional summary information. An excellent example of the type of information available through progress maps can be

found in a report on Australia's National School English Literacy Survey (Masters & Forster, 1997). This report uses maps to display levels of student achievement in writing, reading, viewing, speaking, and listening skills. The level definitions are based on the analysis of empirical data from portfolios of written work from a nationally representative sample of students in Australia, in grades 3 and 5. The map from this study is shown in Figure 4-5. Each of the levels in the map on the left is described by skills that are typical of a student performing at that level and that range from the easiest to the hardest (going "up" the map) for a child to master. For example, of the language indicators on the writing scale, the easiest skills include "Uses some correct initial letters and other sounds" and "Can be read back by the child at the time of writing." The most difficult skills include "Experiments with rearranging sentences" and "Revises writing to be consistent in content and style."

Such a map can be used for a variety of purposes, including summarizing the average and range of performance of students at each grade level, and investigating the differences between subgroups. Because the numerical averages and ranges for groups of students correspond to regions on the map, which in turn are defined by skills typical of those regions, this gives the differences between these groups a substantive interpretation. This is illustrated in the map on the right, which shows the distributions of students in grades 3 and 5 in terms of their locations on the map. For example, in the 2-year span between grade 3 and grade 5, the average performance of students increases from just above level 2 performance, at which they had mastered such skills as "Uses simple sentences" and "Uses repetitive sentence structure," to the upper regions of level 3, at which they were mastering such skills as "Controls simple sentence structure and attempts more complex structures." This sort of interpretation has been used in a number of other assessment systems, such as NAEP, for example, where the levels are referred to as "achievement levels."<sup>8</sup>

Figure 4-6 shows a map from the GSE in economics. This map more closely resembles the traditional item and person map used in item-response modeling. On the left-hand side under the heading of "Persons," we can see a "side-oriented" histogram showing the relative distribution of student estimated locations on the test. To the left of the histogram are two metrics for that distribution—one that gives an interval scale for the measurement,<sup>9</sup> and to its right the percentiles. For the test represented in this map, there were 50 multiple-choice items and a single written-response item scored on a scale of 1 to 5—these are represented on the right-hand side. The multiple-choice items are related to five "strands" or important topic areas within economics, represented by the columns under these headings: fundamental concepts, microeconomics, macroeconomics, comparative systems, and international economics. Nevertheless, the logit scale for reporting student estimates is unidimensional. In addition, these items were designed to represent three different processes or areas of thought emphasized in the economics curriculum: knowledge (K), application (A), and synthesis (S).

From this representation, a number of things can be learned about this examination. For example, it appears that the items on the comparative systems and international strands are on average somewhat easier than items on the other three strands.

<sup>8</sup> See [https://nces.ed.gov/nationsreportcard/guides/scores\\_achv.aspx](https://nces.ed.gov/nationsreportcard/guides/scores_achv.aspx).

<sup>9</sup> The unit for this scale is the logarithm of the odds (logits).

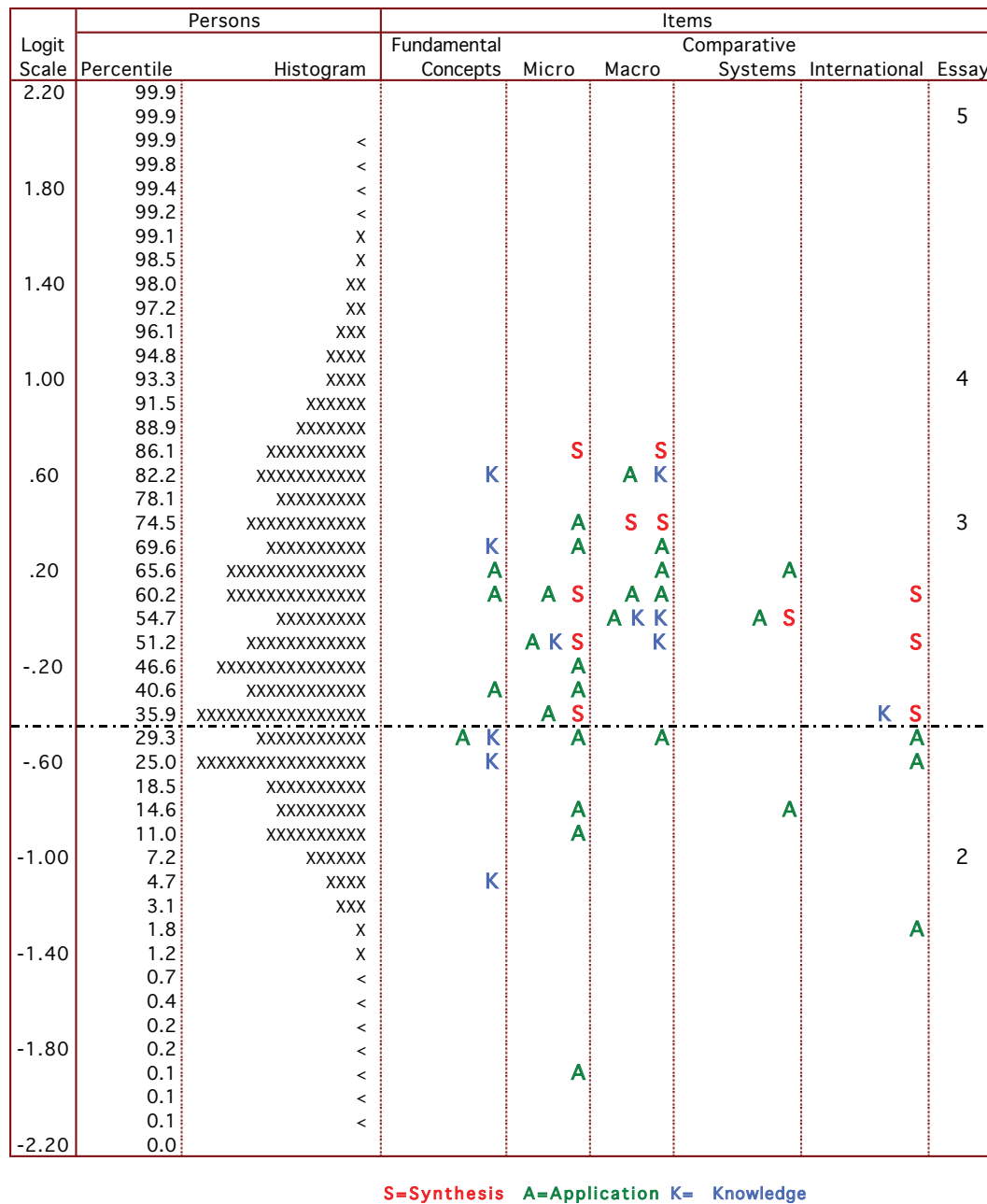


FIGURE 4-6 Wright map of economics GSE by strand.



In addition, within each strand, the synthesis items tend to be on the difficult side of the scale; no synthesis item falls below the horizontal line drawn at approximately -0.5 logits, although a number of knowledge and application items are below this line. Item developers can examine item performance in this way to determine whether items representing the varying strands and processes are performing in accordance with expectations.

Another important thing to note that is made clear by the Wright map is that proximate information about the upper end of the scale is primarily being provided by the levels of the written-response item; the multiple-choice items cluster around the middle of the person distribution, and none appear above about 0.75 logits, or around the 86th percentile. This is especially important because the GSEs were *honors* examinations. The upper three performance levels are the ones for which students receive commendation; being in the lower three performance levels has few, if any, consequences. And in general, only around 30 percent of students tend to fall into one of these top three performance levels; often only 5 percent fall in the “high honors” level. Thus, this representation shows one reason why it is important to have open-ended response items on examinations such as these—they provide information about parts of the scale that are not well measured by multiple-choice items. This may seem an obvious point when presented this way, but that it is not obvious in practice is attested to by the fact that the California State Board of Education closed the GSE program and authorized instead “standards-based” tests composed only of multiple-choice items, even though the new standards-based tests were supposed to measure “high standards.” This sort of policy level decision makes clear that a deeper understanding about what comparability might mean must be developed within the policy maker audiences. The discussion of “stringency” in the next section is directed at just this point.

This section of the chapter has focused on what we see as a bedrock issue for comparability. In the previous section, we asked about comparability in terms of the “content coverage” of the tests in the system. In this section, the focus is sharper: What are the tests actually measuring? Comparability only makes sense if the tests in the system are indeed measuring what is well defined and strongly related to the intended construct. We see that, absent such a quality, the question of comparability can descend to a simple act of “matching numbers” as in a concordance table. What we have attempted to illustrate is the wealth of interpretation that is possible beyond the pale of numerical equivalence. In our view, without this interpretational richness, the comparison of assessments is an exercise in an empty numbers game. However, there is one issue that needs to be clarified before we move on, and that is addressed in the next subsection.

### *Dimensionality Issues*

More generally, the monotonic structure of a progress variable may not be an adequate representation of the outcomes from the curriculum. Instead, a multidimensional framework may be more appropriate, as represented in Figure 4-7. In this example, a two-dimensional construct map for a reading curriculum (Dray, Brown, Diakow, Lee, & Wilson, 2019) is shown, where the two dimensions distinguish between the apprehension of the structures in the text and the understanding about the ideas in the text. Clearly, in such a situation separate Wright maps can be developed for each dimension, and student achievement can be reported as a profile rather than as a single score. One

Level	Construct Map: Use of text structure and how it is associated with knowledge creation	
5	<b>Applying and evaluating communication</b> (and its relation to intent, structure, and features) <ul style="list-style-type: none"> <li>• range of possible choices of structure and features</li> <li>• effectiveness of choices in a particular text</li> </ul>	<b>Synthesizing knowledge and creating new understandings</b> <ul style="list-style-type: none"> <li>• integrating prior knowledge</li> <li>• new understanding based upon multiple texts</li> <li>• evaluating author's intent</li> <li>• literary and/or rhetorical criticism</li> </ul>
4	<b>Judging authorial intent</b> (and how intent relates to structure and features) <ul style="list-style-type: none"> <li>• impact of vocab on reader</li> <li>• impact of structure/features</li> <li>• persuasion</li> <li>• rhetoric</li> </ul>	<b>Cross-checking and coordinating key ideas in the text</b> (seeing the big picture in the whole text) <ul style="list-style-type: none"> <li>• claim</li> <li>• argument</li> <li>• theme</li> <li>• identifying author's intent</li> </ul>
3	<b>Interpreting the structure of a text</b> (and its relation to features) <ul style="list-style-type: none"> <li>• feature hierarchy</li> <li>• foreshadowing at beginning</li> <li>• repeated motifs</li> </ul>	<b>Discriminating key ideas</b> <ul style="list-style-type: none"> <li>• idea structure</li> <li>• supporting statement</li> <li>• plot</li> <li>• characterization</li> </ul>
2	<b>Identifying features of the text</b> (and their relation to surface appearance) <ul style="list-style-type: none"> <li>• text features</li> <li>• language features</li> </ul>	<b>Engaging with ideas in the text</b> <ul style="list-style-type: none"> <li>• topic</li> <li>• main idea of a paragraph</li> </ul>
1	<b>Reacting to surface appearance</b> <ul style="list-style-type: none"> <li>• prominent pictures</li> <li>• title</li> </ul>	<b>Disengaging</b> <ul style="list-style-type: none"> <li>• not challenging existing knowledge</li> <li>• no new ideas</li> </ul>

FIGURE 4-7 A representation of a two-dimensional learning progression for reading comprehension.

could repeat the process described above for each of the dimensions of the framework. This will result in a profile of outcomes for each student, class, etc., including separate sets of cut scores within each dimension, and this may indeed be the desired solution in some circumstances.

However, in many circumstances, although this may well give a more informative representation of student achievement, administrative needs will determine that there must also be an outcome that reaches across the dimensions and gives an overall result. But, resolution of these into a single set of cut points for the overall outcome is not automatically achieved, as in general the cut points in different dimensions will not align in a consistent way, so some additional technique must be developed.

To deal with this, one possibility is to construct a single outcome dimension combining the results from each of the dimensions, and this can be done in several ways. For example, a *reference* dimension (Ackerman, 1988, 1992; Wang, 1986) can be estimated as the unidimensional outcome across all of the individual items across all of the dimensions. A second would be to use a *testlet* approach (Wainer & Kiely, 1987) based on the bifactor model (Holzinger & Swineford, 1939; Schmid & Leiman, 1957) treating each of the original dimensions as testlets. A third would be to use a *composite* model approach, incorporating judged weights among the dimensions (Wilson & Gochyyev, 2020). This overall outcome dimension can then be utilized in a construct-mapping procedure, following through from the construct maps for each dimension to the outcome dimension. The suitability of these different modeling procedures can be readily assessed using standard model-comparison techniques. The reference dimension and testlet approach have somewhat higher hurdles, as they require that the outcome dimension be found to be reasonably unidimensional, whereas the composite dimension does not make such a stipulation.

The requirement of unidimensionality across the grades is one that is likely to cause problems when the span of grade levels is large. This should be clear from examination of typical subject-matter content over the grade span. For example, early elementary grade mathematics will feature basic arithmetic, while upper elementary grades tend to focus on algorithmic manipulations, and this will change to a focus of high school algebra in middle school. While these might all be labeled as “math” in reporting to parents, they are actually quite different constructs and are themselves composed of different sets of subcomponents, and this will make the maintenance of a consistent scale across these distinctions quite challenging. The use of a nondimensional approach such as the composite one mentioned above will lessen this technical problem, but still an interpretational issue will remain, and hence the reliance on the vertical scale may need to be delimited over longer grade spans.

### COMPARABILITY OF STRINGENCY OF ASSESSMENTS BETWEEN SUBJECT AREAS AND BETWEEN GRADES

*Stringency* in an educational achievement test refers to whether high scores are “appropriately” rare. This is difficult to define without circularity:

1. If a test is easy, there will be more high scores, and if a test is difficult, there will be fewer.

2. If students have high achievement, there will be more high scores, and if achievement is low, there will be fewer high scores.

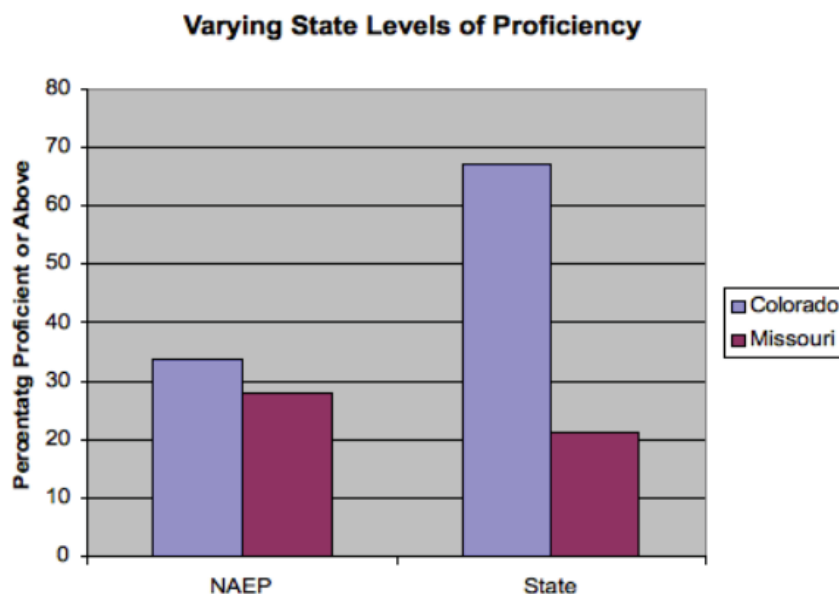
But any test scale can be translated up or down or multiplied by a constant, so these relationships of student achievement can be arbitrarily manipulated. The same is true of performance levels defined by cut scores on the scale. This issue of stringency can arise in several different contexts. One common context is where comparison is to be made between two different assessments of the same subject conducted in two different states. This is a topic that is discussed in Chapter 5, *Comparability Across Different Assessment Systems*. However, we use it here as a somewhat more familiar context in order to introduce the concept of stringency.

An example using NAEP may help clarify this issue. Currently, different states (or sets of states) develop their own tests to assess the students in their educational system. One question is whether we can compare across these different tests to see if some states are doing better than others. The comparison must be made in such a way that the state tests are adjusted for their stringency. In this situation, there is indeed a national standard that can be used to evaluate the stringency of state tests—NAEP. Thus, for example, Figure 4-8 shows a comparison between two states, Colorado and Missouri (Linn, 2005).<sup>10</sup> The right-hand panel shows that, according to their respective state tests, students in grade 4 in Colorado are doing much better than students in Missouri in the topics of reading and language arts—in fact 67 percent of Colorado students were judged to be at “proficient or above” while only 21 percent of Missouri students were judged so. However, the left-hand panel of Figure 4-8 shows that, according to the national test, NAEP, these percentages should be much closer: more like 34 percent for Colorado versus 28 percent for Missouri. We can see that the Colorado test was much less stringent than the Missouri test, and that the NAEP test was somewhat less stringent than the Missouri test, and quite a lot more stringent than the Colorado test. Note that this question of stringency would best be asked conditional on a finding that the content of the two tests was reasonably consistent (as discussed above in the section “Defining the Content of an Assessment System”); however, this question is frequently ignored in such comparisons.

The contexts that we are concerned with in this section are different from the state-to-state comparison illustrated in the previous paragraph. Both of these contexts occur within an assessment system: The first is the comparison of two tests of the same subject matter across two different grades (an example would be the comparison of a grade 5 mathematics test with the grade 6 equivalent). The second is the comparison of two tests on different subject matters at the same grade level (an example would be the comparison of a grade 5 mathematics test with a grade 5 reading/language arts test). Now, it may be difficult to ascertain the stringency of a specific test for a specific population (e.g., grade 4 reading/language arts), but the problems are necessarily confounded in these two cases, which involve multiple populations and multiple tests.

---

<sup>10</sup> Note that much more extensive examples are available at <https://nces.ed.gov/nationsreportcard/studies/statemapping>.



**FIGURE 4-8** Relationship between state proficiency levels and scores on NAEP.  
SOURCE: Adapted from Linn (2005).

### Comparison of Tests of the Same Subject Matter Across Different Grades

The question of comparing across grades is one that is germane to any assessment system—educational stakeholders want to know that students are increasing in their achievement as they move from grade to grade, and this cannot be accurately ascertained unless there is a means of comparing students across the tests that are used at different grade levels. The broad expectation is that, as students move up in grades, the test should be different, but in a specific way.

One important thing to bear in mind when considering the design of tests intended to measure the same subject matter across grades is that the question of having the same content, as examined earlier in this chapter, does not directly apply. Instead the relationship between the contents of the tests must be seen as being more complex—they should be similar enough to deserve to be placed in the same subject matter, but they should be systematically different, as appropriate for tests that occur at different points on a student's school learning; that is, they *should* differ in difficulty.

Differences in test difficulty across different grades cannot be examined from *within* the perspectives of the samples of students at the different grade levels. Normative information about student performance within each grade, such as median scores for each grade on its own test, conveys no information about the relative difficulties of the tests. What is needed is a specially designed sample of test takers who take items common to the two grade-level tests. A very special arrangement would be for a single sample of students to take the tests for all of the grades. Then one could express, for example, the grade medians as percentiles on this amalgamated test, and this could be carried through to finding percentiles for performance levels and so forth. Unfortunately for this possibility, the practicalities of testing mitigate against having students



take a large number of tests, especially tests that contain items far below or far above their range of expected success. This would also cause technical difficulties with the scaling. Hence, one is forced to develop alternative solutions, which will of necessity involve a more complex strategy that makes it difficult to use simple normative or classical true score approaches.

It is beyond the scope of this chapter to explain in detail the several methods that have been developed to address this challenge. Instead we constrain our discussion to a strategy based on the approach described in the section “Design of the Constructs in an Assessment System,” taking advantage of its strengths to make observations about the issues raised above regarding stringency. The general approach we adopt is that of constructing a vertical scale (Briggs, 2013; Kolen & Brennan, 1995) that spans across the grade levels of the assessment system (or, at least, some substantial part of the system) and utilizing that to examine, and interpret, the relative stringency of items from the different grade levels.

Of course, as outlined earlier in the chapter, prior to the construction of a vertical scale spanning grade levels, a construct map that covers the range of the relevant grades should be developed. An example could be derived from the construct framework for algebra shown in Figure 4-4, or for writing achievement shown on the left-hand side of Figure 4-5. Assuming that this has been accomplished, then what is required is to develop a collection of items that span the same range as the entire construct map, not just the content for each grade, and to test progressive samples of students that take overlapping sets of items from the lower to the higher grades. There are many issues and details about how these samples are defined and how the data analysis takes place, but the central concept is that there is linkage in the data between overlapping sets of students who take common items across grade-level tests, and items that are taken by sets of students across grades. Quality control needs to be applied on the resulting item sets, just as it does for grade-level tests, but the results can be displayed in Wright maps similar to that in Figure 4-6, and the grade-level results can be distinguished as shown in the right-hand panel of Figure 4-5.

Armed with such a vertical scale, the issue of stringency across grade levels becomes a matter of judging an appropriate cut score for a particular level on this vertical scale, based on (1) the content definition of the construct, (2) the empirical information obtained from the special sample about the relative difficulties of the items at the different levels (i.e., as expressed in the Wright map), and (3) the impact data regarding how many students would be included in each category, as obtained from the large-scale samples (or censuses) at each grade level. This last will involve some attention to score distributions on the vertical scale. For example, if obtaining consistent percentages proficient across grades required setting successive cut points at very different locations in the raw-score cumulative distribution functions for successive grades, then the analyses and results would need to be checked. The focus in a standards-based application of this approach would be the specification of cut scores for performance levels such as “basic” or “proficient” for a succession of grades. Again, there are several ways to address this, but we only describe one approach in order to illustrate the steps and enlighten the issues involved. We again use the Golden State Exam context to illustrate this approach, which was indeed designed to be used across grade levels. The challenge that must be addressed is to respond to the criticism above (in the section



“Defining the Content of an Assessment System”) about the “mystifying” labels that are typically used for performance levels.

A crucial use of Wright maps in the GSE program was for setting the cut scores between the six performance levels (such as those shown in Figure 4-4).<sup>11</sup> The “construct mapping” method (Wilson & Draney, 2002) was developed to allow the members of a standard-setting committee to use their knowledge of the curriculum and their understanding about how the construct is made manifest through the items as a basis to model what a student at a given level knows and can do, and thus to map their judgments onto the Wright map response. Dichotomous and polytomous items are scaled together to produce difficulty thresholds for the items and proficiency estimates for the students. The calibration is then used to create a Wright map combining the locations of the multiple-choice items and the score levels of all the performance items. The map is represented in a piece of software (Hoskens & Wilson, 1999) that allows committee members to determine the details of student performance at any given proficiency level, and to assist them in deciding where the cutoffs between performance levels should be.

An example showing a section of such a Wright item map is given in Figure 4-9. The column on the far left contains a numerical scale that allows the selection and examination of a given point on the map and the selection of the eventual cut scores for the performance levels. This scale is a transformation of the original scale, designed to have a mean of 500, and to range from approximately 0 to 1,000. The next two columns contain the location of the multiple-choice items (labeled by number of appearance on the examination) and the probability that a person at a selected point would get each item correct (in this case, a person at 500 on the GSE scale, represented by the shaded band across the map). Note that this point can be changed by the user to check the results of setting the cut score at different levels. The next two sets of columns display the thresholds for the two written-response items. For example, the threshold levels for scores of 2 and 3 on written-response item 1 are represented by 1.2 and 1.3, respectively (although each item is scored on a scale of 1 to 5 on this particular examination, only the part of the scale where a person would be most likely to get a score of 2 or 3 on either item is shown). The second set shows the probability that a person at 500 on the GSE scale would score at that particular score level on each item. The software also displays (not shown in Figure 4-9), for a person at the selected point on the GSE scale, the expected score total on the multiple-choice section and the expected score on each of the written-response items. The software also allows viewing of item content and, for open-ended items, scoring guides and exemplars as discussed above.

In order to set the cut points, the committee first acquaints itself with the test materials. The meanings of the various parts of the map are then explained, and the committee members and the operators of the program spend time with the software, familiarizing themselves with interpretations of different points on the scale.

The display of multiple-choice-item locations in ascending difficulty, next to the written-response thresholds, helps to characterize the scale in terms of what increasing proficiency “looks like” in the pool of test takers. For example, if a committee was

---

<sup>11</sup> Note that there are several other methods that are available for the setting of cut scores. See Cizek and Bunch (2006) for a description of many of them. However, only some of these methods would make use of the construct or analogous content conceptualizations.

GSE Scale	Multiple Choice		WR 1		WR 2	
		P		P		P
620						
610						
600						
590						
580						
570	37	.30			2.3	.26
560	15	.34				
550						
540	28 39	.38				
530	27	.41				
520	19 38	.45				
510						
500	34 43 45 48	.50	1.3	.40		
490	17 18 20 40 50	.53				
480	4 31	.56				
470	11 32 33 44 47	.59				
460	5 9 12 46	.61				
450	3 6 7 10 16 29	.64				
440	36	.67				
430	8 14 22 23 26 35	.69				
420	13 24 25	.71				
410	41 42	.73				
400	1 21 30 49	.76				
390						
380					2.2	.56
370	2	.82				
360			1.2	.40		
350						

FIGURE 4-9 GSE cut-point-setting map.

considering 500 as a cut point between performance levels, committee members could note that it is a point at which items like 34, 43, 45, and 48 are expected to be chosen correctly about 50 percent of the time, a harder item like 37 is expected to be chosen correctly about 30 percent of the time, and easier items like 2 are expected to be chosen correctly 80 percent of the time. The multiple-choice items near any chosen point can be seen by the committee so the members can relate these probabilities to their understanding of the items. The committee could also note that a student at that point (i.e., 500) would be equally likely to score a 2 or a 3 on the first written-response item (40 percent each) and more likely to score a 2 than a 3 on the second (56 versus 26 percent). Examples of student work at these levels are also available to the committee for consideration of the interpretation of these scores. Committee members can examine the responses of selected examinees to both the multiple-choice and written-response items, chart their location on the map, and judge the level.

The committee then, through a consensus-building process, sets up cut points on this map, using the item-response calibrations to give interpretability in terms of predicted responses to both multiple-choice and open-ended items. Locations of students on the scaled variable are also available for interpretative purposes. This procedure allows criterion-referenced interpretations of cut scores as well as the traditional norm-referenced interpretations. The focus in a standards-based application of this approach would focus on cut scores for performance levels such as “basic” or “proficient” for a succession of grades.

Use of the maps available from the item-response modeling approach allows the committee to interpret cutoffs not only in a norm-referenced way, but also in a criterion-referenced way, as described above. This can then be used as the basis for checking the consistency with curriculum documents, and for reporting back to teachers and other education professionals in terms of performance levels such as “basic” or “proficient” for a succession of grades. Another crucial check is the stability of the cut scores in terms of the percentage that are in each subgroup (e.g., percent “proficient”) across successive grade levels. There is often a need for some “smoothing” so that the percent proficient level does not jump around. Otherwise, misinterpretation of capricious cut scores may lead to inappropriate reallocation of resources toward “underperforming” grade levels.

On the technical side, the underlying response scale must be rechecked for consistency. One immediate check would be to reserve a fraction of the data samples on a representative basis for cross-validation of the statistical results including item and person parameter estimates, as well as the conclusions from fit analyses, and also for reporting of the relevant outcome results (such as mean differences between grade cohorts, etc.). Over time, care would need to be taken to check for parameter drift and for substantive changes in outcome results.

### **Comparison of Tests on Different Subject Matters at the Same Grade Level**

Another context where the concept of stringency is essential is the comparison of tests on different subject matters at the same grade level. For example, is a grade 5 mathematics test more stringent than a grade 5 language test just because there are fewer high scores in mathematics? Or does that mean students are better in language

than in mathematics? Or did the language test developers create “easier” items than the mathematics test developers?

Very often, the reporting scales for two tests for two different subjects, such as mathematics and language arts, are reported in ways that are numerically similar. See, for example, Figure 4-10, which is a summary of the scores for the performance levels for the State of California performance levels in mathematics and English language arts (ELA) (California Department of Education, 2019, p. 56). Also, the labels of the performance levels and even their description may be couched in quite general language that looks very similar across tests (e.g., see the labels in Figure 4-10). If the labels are not tied to concrete performances, then it is unclear what they might mean. This potentially leads to a reification of the labels and score interpretations, as well as some questions: Is this usage of labels such as “standard met” just a matter of convenience or is it intended to deliver a message? Or is it deceptive?

Grade Five Scale Score Range for ELA and Mathematics

Achievement Levels	Level 1: Standard Not Met	Level 2: Standard Nearly Met	Level 3: Standard Met	Level 4: Standard Exceeded
ELA Scale Score Ranges	2201–2441	2442–2501	2502–2581	2582–2701
Math Scale Score Ranges	2219–2454	2455–2527	2528–2578	2579–2700

**FIGURE 4-10** Scores for the performance levels for the state of California performance levels in mathematics and English language arts (ELA).

The learning progression perspectives described in the section “Design of the Constructs in an Assessment System” and in the paragraphs above offer some perspectives on comprehending the stringency of different tests of the same subject and across grades. And these perspectives may be applicable in the context where different subject matters are involved. Indeed, the composite model could be applied to establish overall outcome variables across different subjects such as mathematics and language arts. The work of the standard-setting committees would likely be more difficult, as teachers and experts may not themselves span across these subjects; hence, the whole committee would have to find ways to bridge between the different views.

One obvious approach to this problem involves norm referencing (E. Haertel, personal communication, August 2019). In fact, a variant of the construct mapping approach, based on normative rather than scaled-score maps, might be useful for judging student performances across tests of different subject matters. The temptation would be to set the percentage, say, *proficient*, at the same point across all dimensions, but this would need to be carefully considered.

If we hope to make a more rigorous definition, analysis, and comparison of stringency, one way would be to introduce some kind of concrete reference. A natural choice would be to refer to the time and effort required to reach a given level of accomplishment. For example, we could refer to the level of proficiency that requires, say, 40 hours of classroom instruction, or 10 hours of tutoring. That approach probably seems impractical to implement, but it may be the essential conceptual justification for dealing with

measurements of achievement in “grade equivalents.” And therefore, it is a justification for taking a serious look at vertical scaling.

## RELIABILITY FOR DIFFERENT USES<sup>12</sup>

### General Principles

If two achievement tests are supposed to be linked<sup>13</sup> and scores compared, over grades or over time as prime examples, then the *accuracy* in the comparison refers to the closeness of the measurements provided by linked tests when repeatedly derived, at least in theory, from the same or equivalent individuals or groups. The deviation from closeness can be systematic and persistent, which is termed *bias*, or it can be random variation or fluctuation, which is termed *precision*. Test comparability can be more or less accurate, in terms of both bias and precision, at different parts of the score scale. In this section, we focus on two major reliability concerns: linking and measurement error. Then we look at reliability requirements, especially in terms of reporting scores and subscores for individuals and groups.

### Linking Error

Accuracy in a technical sense is conceived of as the result of statistical replications of the linkage, using different samples of items or respondents. The closeness on the average over replications of the comparison to the true or reference comparison is the trueness (nonbias) and the variance over the replications is the precision.<sup>14</sup> One way to define the true comparison is to consider a very long test and very large student sample.

Bias can be due to an imperfect or inadequate selection or positioning of items used for linking and their interaction with the characteristics of the populations being tested. Bias also enters through imperfect statistical models for linking and through interactions with the other, nonlinking items used in the analysis. Of course, we cannot actually measure bias; in that case, we would simply remove it.

Precision in linking is the potential variation or fluctuation in linking and is an effect of the particular items and student samplings used in the linking analysis; it can arise from the use of different linking items, different nonlinking items, and/or different linking student samples. Precision is a variance that can be estimated but not corrected.

---

<sup>12</sup> This section on reliability and its connection to assessment design and reporting uses statistical and psychometric terminology and technology for educational-psychological testing. The classical perspective is articulated in Cronbach (1990) and the item response theory perspective in Embretson and Reise (2000). The general procedures and details of equating can be obtained from Kolen and Brennan (2014). The approach taken in considering components of true and error variance between students and aggregates and across subdomains comes from the theory of generalizability of Cronbach, Gleser, Harinder, and Rajaratnam (1972). Integration of these matters is found in Cronbach, Bradburn, and Horvitz (1995).

<sup>13</sup> The term *equating* is used when the scales of two tests constructed from the same blueprint are aligned for the same or parallel test-taker populations. The term *linking* is more general and used as well when other approaches for aligning scales for tests that are partly overlapping or populations are different. See Mislevy (1992) for fundamental issues and Kolen and Brennan (1995) for detailed procedures.

<sup>14</sup> These are definitions from the International Organization for Standardization (ISO) 5725-1 standards (ISO, 1994).

Linking error of both kinds, bias and precision, is an underappreciated part of assessment systems. It occurs whenever statistical and psychometric efforts are made to align the scales and scores of two tests, especially between years at one grade but also in vertical aligning over grades in one year. Linking error surfaces in the eventual scoring of student tests and in student and aggregate (class, school, district, and state) reports. We refer to this as *wobble*; that is, from year to year, achievement results will seem to rise and fall by some amount that is not really an interpretable trend—it is just the unavoidable consequence of imperfect linking. Similarly, there will be wobble within and across years in the between-grade differences and trends. We cannot remove wobble, but to make valid inferences about testing trends, we need to estimate its magnitude, as a standard error, and take care that the wobble is smaller than the meaningful stabilities and trends in the achievement.

Another source of wobble arises when test score distributions are discrete, with score points lumped at a relatively small number of values (e.g., corresponding to identical response patterns or whole-score totals). Then alignment of linked tests may lead to substantial differences in results such as percent above criterion, creating wobble that might suggest policy inference but is really due to granularity in the reporting scale.

Real changes in achievement occur because of real factors of teaching and learning, instructional time, population composition, and, of course, student learning. We hope to measure *growth*—individual (or aggregated) student change in knowledge, skills, and abilities within a grade or from one grade to the next. And we also hope to measure *trend*—performance of successive annual cohorts at a given grade level. There should also be interest in the trend of growth. These kinds of changes are not linking errors but reflect meaningful change in individual and population achievement. If we remove those kinds of trends by standardizations (such as percentiles, etc.), we would be distorting the realities of achievement. Differences in testing conditions and in student test-taking motivation need to be considered as well, and perhaps these are controllable or can be influenced; they are not what we would ordinarily consider real trend or uninterpretable wobble.

Linking obviously depends on the exact equivalence of the linking items between the administrations of different test forms. Changes in the presentation of the items, including typography, size and position of pictures and diagrams, placement on the page, and order of alternatives, can easily create shifts in difficulty and interfere with linking. Even if items on two forms of a test (such as for successive years) seem to be identical, there may be other factors, including position and context (i.e., nature of the surrounding items), that can create confusion in the linking. More generally, when calibration using common items is carried out, there may be effects of multidimensionality of the linking item set. Finally, random estimation error creates some wobble due to item and student sampling.

### Measurement Error

The test score for an individual student includes another potential for inaccuracy: In the process of answering a test a student naturally fluctuates in attention, has or fails momentarily to have insight into the question being posed, guesses, responds carelessly, mistakes the response method, etc. These can be considered random acts of precision



or imprecision in the response process. More fundamentally, a student's knowledge and skill match up idiosyncratically with the demands of the test items. Items that are generally easy may be especially hard for some students, for example, if they missed the corresponding instruction. Items that are generally difficult may be especially easy for some students, for example, if they have recently received instruction in the area or if they have particular interest in the specific content or context. Both random variation in the response process and interaction between student and item that might be persistent create, from the perspective of scoring and measurement, item-by-student interaction variance, and this is termed *measurement error*. Change scores get a double dose of measurement error.

Measurement errors can also be salient in aggregate results. Another underappreciated aspect of achievement accuracy is the presence of group-by-item interactions (e.g., class, school, student group, or district). These would derive from group differences in instruction and curriculum (detailed above in the "Opportunity to Learn" section). A common problem even when an established curriculum is similar is that the pace or timing of instruction has different classrooms and schools being tested when different parts of the curriculum have been covered. The sizes of the groups also affect the precision of the results. This is obvious when different students take different item samples, in matrix test administrations or CAT. But even when there is a single form involved in aggregation, there will be effective sampling error when the inference is to the characteristics of the aggregate group, say, from one year to the next. These issues were discussed by Cronbach, Bradburn, and Horvitz (1995) in evaluating the California Learning Assessment System, which collapsed, in part, because of excessive measurement errors.

For someone constructing a test or working on a component of testing and scoring, for example, the scoring in a subdomain or the scoring of constructed responses, the determination of "reliability coefficients" is a useful tool. They are defined as the proportion of the variance of the total score (or estimate) that is true—not error. For someone interpreting or using a test score or its aggregate summary (mean), the better index of quality is the *standard error of measurement* (SEM), which speaks directly to the precision of the result. Importantly, it is in the metric of the test scores and can be used to estimate statistical confidence intervals or to determine the significance of score or mean differences. Often one sees claims about the accuracy of test scores based on internal and partial reliability coefficients. For example, the accuracy of the scoring of student writing is said to be high if the correlation of scores from two judges is high. While scoring and interrater reliability contribute to measurement error, the contribution is not straightforward: First of all, there is variance between judges, variance between students, and interactions between judges and students. Second, the correlation, or reliability, does not translate directly into the precision of a student's score on the writing or into the precision of the final report, which is presumably a composite of the writing score and other items. This logic extends to aggregate results, such as classroom, group, school, or state means. The process of estimating reliabilities is important in building scoring and reporting systems, but we need to estimate the bottom-line accuracy of the results and the SEM is better for that.

### **Accuracy Requirements: Total and Subscores, Individuals and Groups**

What are the accuracy requirements for individual measurement, for aggregate group summaries (schools, districts, subpopulations, etc.), and for monitoring change over time? The standard errors of test scores and aggregates impose important limitations on our ability to answer policy questions and limit the impact the results should have on decisions that might be made about individuals, schools, programs, and so on. In answering these questions, it is important to have honest and complete information about bottom-line accuracy, including the influences of (1) measurement errors due to item sampling, item-by-person interactions, rater variance, and student response error (e.g., from inattention and guessing) and (2) linking errors. For individual student scores, the regular measurement errors will typically swamp the linking errors. But in aggregate results, such as class or school means, the individual measurement errors will tend to cancel out and have reduced impact, and hence the linking errors become relatively more important, and aggregate-by-item interactions (class, school, or district) will persist as well.

The simple but essential definition of the accuracy that is required for individual or aggregate scores is that it must be sufficient to justify the inferences that will be made from the scores. If there is to be a decision about a student (pass/fail, selection for remediation or for enrichment, or ranking) or if the score will be taken to indicate the quality of the student's knowledge and skill, then the confidence interval for the score, constructed from the standard error, should contain only values consistent with those interpretations and uses. The same logic applies to accuracy requirements for aggregate scores, considering the realm of decisions and interpretations that will be applied for classes, schools, districts, and so on. Specific determination of accuracy requirements depends on the educational and policy applications, but it also depends on knowing and applying correct, all-inclusive standard errors.

Probably the most important and controllable source of linking and measurement error derives from the sample of items in a test. From the perspective of other educational activities—such as classroom instruction or student project work—conventional large-scale tests are short and not very well focused on particular specific knowledge and skills. This remark should not be so surprising, as, after all, the typical large-scale achievement test must measure students across a very wide range of abilities and concepts. (Adaptive tests may do a better job of focusing, but usually along just one dimension.) Of the many thousands of potential items or lines of inquiry that could be used and that give detailed coverage of a content domain, we choose perhaps 50 or 100 items, too often chosen for statistical correlation to a central dimension and not to cover all corners of the domain.

How does this affect the design of the assessments and the quality of the results? Fundamentally, it means that a test is a very small sample of the domain. If we are to equate two tests with mostly different item samples, and the domain has any amount of internal variability, then it will be impossible to ensure a precise equating: wobble is inevitable. Again, this is probably of less importance for the accuracy of individual total scores and reports, but it will be critically important for aggregate results and for trend.

The problem of item sampling is much more severe if results are needed for subdomains, such as different topics in mathematics or science or different skills in reading. Then the item samples are vanishingly small. In the example that follows (see Figure 4-11), 100 items are used in a domain and there are five subdomains, so there are only

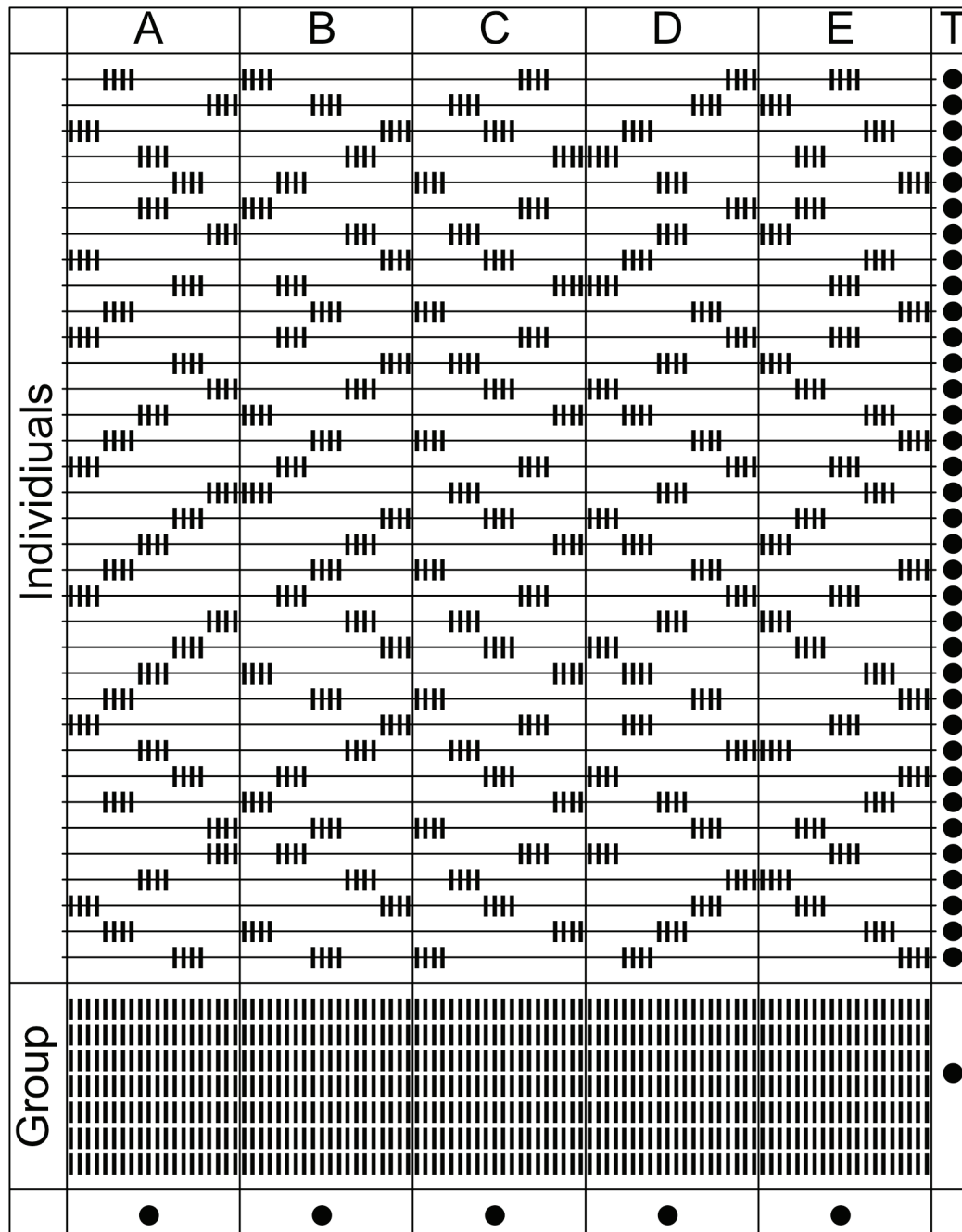


FIGURE 4-11 Illustration of stratified multiple matrix item sample design.

20 items per subdomain.<sup>15</sup> It is conceivable that a meaningful score can be obtained from 20 items, but it is difficult to calibrate and equate them from one test to another or over time. We should have a rule that subscores should not be reported unless we can ensure their comparability between tests and over time. We should not report subscores to individual students except with the same seriousness that we report total scores. However, that approach is not feasible unless tests are much longer. Adaptive testing might provide some relief, but it will have to have multiple targets.

Designs exist that obtain adequate accuracy for individual students along with much more detailed information for groups, but this is only possible if we can overcome simplistic rules about everyone taking exactly the same test, and simplistic methods for aggregation, in which individual scores or subscores are averaged by class, group, and so on. One kind of design using item-sampling methods is *multiple matrix sampling*. This is illustrated in Figure 4-11. As mentioned, this shows a test with 5 subtests of 20 items. The dots on the right-hand side show total score estimates for individuals. The dots along the bottom show subdomain scores for a group (classroom).

The best reports will likely be complex statistical estimates, perhaps including demographic information as well (as is done in NAEP), and not just simple observed scores. An early attempt at extending these approaches was the “duplex design” (Bejar & Graf, 2010; Bock & Mislevy, 1981). CATs as used in SBAC and PARCC have the potential for providing broad representation of the content domain, but very complex methods (quotas and balancing) would be required to ensure subdomain scores even for groups.

### General Advice

In considering these points made above, and reflecting on experiences of working with state and national testing programs, we offer the following precepts: when developing tests for future assessments, we should

1. Realize, and help our political and public colleagues to understand, that having different tests for different students, with matrix sampling or adaptive testing, can provide fair and accurate testing for individual students and do much more.
2. Promote test designs that have adequate accuracy for individuals along with much more detailed (subtest) information for groups. To do this, we need to use matrix sampling or CAT and get away from simplistic aggregations. This will require complicated analyses and explanations.
3. Make more comprehensive error analysis that admits item- and subtest-by-group interactions as contributors to measurement and linking error.
4. Improve the quality control in testing and analysis so that we detect and correct data collection errors, item problems, and scaling irregularities.
5. Understand that the design and analysis of assessment systems is complicated and at times highly technical; this requires expert advice and oversight, and

---

<sup>15</sup> Unfortunately, in many assessments, there are more subdomains to be measured and fewer items available to accomplish that.

contractors should be held to high standards of analysis and transparency by an oversight group.

### TRANSPARENCY IN THE DESIGN AND INTERPRETATION

Providing clear and readily available information about an assessment system is seldom a highlight in the proposals for system development and is often just an afterthought in the actual implementation. However, the provision of such is in fact crucial for the success of the system, and for the believability in the operational integrity of the system. We discuss two components of this transparency, in turn:

1. What is the quality of the documentation that is available to the *consumers* of the reports from the system that will help them make appropriate interpretations of the information contained in the reports?
2. What is the quality of the documentation that is available to the *evaluators* of the system that will help them judge its accuracy and effectiveness, as well as suggest improvements?

#### For Consumers

First, the success of an assessment system should be judged on two bases: (1) the quality and usefulness of the outcome results that the system reports, at the various levels of intended use (i.e., does it fulfill its *informational* purposes?), and (2) the quality and usefulness of the system beyond the outcome results (i.e., does it fulfill its *signification* purposes?<sup>16</sup>) (Wilson, 2018). Issues of comparability are not directly engaged with the latter purposes (important as they are); hence, we focus on the former. The essential question here is whether the system gives sufficient training and support materials to at least make it likely that consumers will make similar (and similarly correct) interpretations of the specific outcomes from the system. We can approach this question by considering the comparability of two consumers trying to interpret the same test, or one consumer trying to interpret equated versions of two tests.

For example, if parent A receives a report that says that their child has attained an outcome of, say, 305, will that interpretation be similar to the interpretation of parent B, whose child's outcome was the same? A normative view can provide one sort of answer to this question. If the estimate of 305 is reported to be at the 50th percentile for students in the grade level (as is the case for year 3 students in Figure 4-5), then this raw fact is equally meaningful for both parents A and B. And, this can be made more specific to particular comparison groups, such as males and females, ethnic groups, etc., by reporting percentile locations for those groups. However, although it may be somewhat comforting for the parents to know that their child is at the median (50th

---

<sup>16</sup> This was in earlier publications referred to as the "threat" form of coherence, as embodied in standards and items included in an assessment-based accountability system (Wilson, 2004). We would see that there is a comparability concern if a test or its documentation does not reveal its real content, which might be a fraction of the labeled content, leading consumers to misunderstand what is being tested or not being tested.

percentile), the interpretability is limited; the parents still have no idea about what it is that their children actually know or do not know. To respond to this, one might consider the domain view discussed in the section “Defining the Content of an Assessment System” above. However, this can lead to considerable misunderstandings, as the most salient connection for an uninformed parent to make would be to assume that being at the 50th percentile meant that students had “mastered” 50 percent of the content of the standards for that grade. But this would be distinctly wrong—the percentiles and the percentages reflect completely different ideas and must not be confused.<sup>17</sup> In a context where these results were sufficiently focused on a construct, then the reports could include support in the shape of Wright maps such as that shown in Figure 4-5. Then the parents could interpret that their children were performing at the level indicated by the descriptions of level 3, such as “Shows some evidence of planning, revising and proof reading own writing” and “Spells many common words correctly.” Supplementing the Wright maps with actual student responses to specific test items can help make these descriptions more concrete (although care must be taken to ensure that readers do not overinterpret the occurrence of specific items and responses, and are clear that these are just samples). In a coarser-grained context (such as a broadly defined test on a whole subject area, like “science”), it is more difficult to develop strong materials to support this sort of deep content-construct interpretation. Nevertheless, broad level descriptions such as those shown above (in Figure 4-4) for algebra may be useful, but also norm-referenced interpretations may be meaningful to consumers. This illustrates how scale outcomes must be reported, along with content and construct-related materials, in order to help maintain comparability of interpretations.

An analogous question can be asked at higher levels of aggregation: If a school has attained a mean score of 200 for grade 3 students (on the scale in Figure 4-5), say, will the interpretation of that score by the school’s principal be the same as that for another principal whose school had the same outcome? Again, one could look to a normative framing for the interpretation, but one would need to be careful to use the correct distribution, as for this question, it is the school percentiles that are most appropriate, not the student percentiles (which will be quite different).<sup>18</sup> But the interpretations will again be limited by the nature of the normative results. In contrast, the construct-oriented interpretation utilizing the Wright map will be equally as informative and straightforward as it was for individual students—average students in grade 3 will be performing at level 2, and then the level 2 description and associated materials can be consulted for interpretations.

Going beyond the interpretation of the scores themselves, one can ask whether the scores *ought* to be compared, due to differences in the situations of the two schools, that is, whether the student bodies served by the two schools are of a comparable makeup (same percentage of English learners, same distribution of parent education levels, same community resources, etc.). Questions like these are not simply measurement

---

<sup>17</sup> Indeed, a domain-based interpretation has been proposed by Bock, Thissen, and Zimowsky (1997), but this has not, so far as we know, been used in any large-scale applications.

<sup>18</sup> Looking at Figure 4-5, the student percentile for an estimate of 200 would be the 21st percentile, but the school percentile would be much lower (as the school distribution will inevitably be narrower than the student distribution).



comparability questions but invoke broader educational policy issues. For instance, consider the issues involved in the usage of the results across contexts where the educational situations differ substantially. Two schools that were found to be at the same estimated location on a certain test (within measurement error) could be said to be performing at about the same level. However, the interpretation of that level of performance may need to be adjusted depending on the enrollment context of the school: a school with substantially more English learners may be judged to have worked better than one with less. Of course, any single demographic difference would need to be considered in concert with other relevant factors. And, this issue is made even more complex when the outcome in question is an index of educational gains, as here the stability of the contexts over time, as well as the issues of vertical scaling, will also be relevant.

A somewhat different question arises when two students with similar but different estimates are compared. For example, if one receives an outcome of, say 400, and a second receives an outcome of 450, should the interpretations be different? Here the question arises as to whether this difference is meaningful at all, and again, just as above, there are two different perspectives. One perspective is a statistical one, associated with the reliability, as discussed in the section “Reliability for Different Uses.” The interpretational materials must include not just the reliabilities but also the standard errors of measurement for each estimate. Suppose that, in the case of Figure 4-5, the standard error of measurement was reported as being 10 for the estimates around 400. Then one could ask the question of whether the two estimates, 400 and 450, could have been generated by a student with the same underlying location on the variable. At a typical 5 percent level, this difference would indeed be considered statistically significant.<sup>19</sup> However, one must also always look beyond statistical significance and ask about the practical significance (i.e., effect size; e.g., Kohler & Hartig, 2017) of this difference. If we posit that the “metric” of the effect is the levels in Figure 4-5 (of course, other metrics might be useful for other purposes), then both estimates remain within level 4, and hence, in contrast, the two scores are not practically significantly different when one is using the “writing levels” as the frame of interpretation. Kane (1996) introduced the very helpful notion of “tolerance for error”: “The tolerance for error specifies how large errors can be before they interfere with the intended use of the measurement procedure” (p. 355). This may provide additional interpretative leverage, although, of course, it would need to be well prepared in advance. As above, this sort of question can be asked at higher levels of aggregation, such as for schools. Here the testing of statistical significance becomes more complicated, as sampling error must be considered as well as measurement error (see footnote 12 in the section “Reliability for Different Uses”). However, the interpretation based on the Wright map is no more complicated than it was for students, as again it directly relates to the writing levels. Because standard errors for school means will be small relative to standard errors for individuals, smaller differences between schools are likely to be significant, and the chances of crossing level boundaries will be lower.

---

<sup>19</sup> That is, for a student located at the midpoint of the two, 420, the 95 percent confidence interval would be (403.4, 446.6).

### For Evaluators

Those responsible for the development and implementation of an assessment system (i.e., the policy makers and funders) will need to ensure that the system includes the sorts of interpretational materials as described in the previous subsection. But, they will need to go beyond to provide to evaluators a comprehensive and transparent documentation of the complete assessment system, including all the processes that they went through in devising the system. This may not obviously be an issue of comparability, but we do indeed make that claim. We see that the highest standard of documentation would be that the system itself could be duplicated (in all its detail and performance) using this documentation, by a second development team who replicated the system. Of course, this is a very high standard, and one that we see as being based on the logic of hypothetical thinking—we do not expect that any evaluator would actually replicate the whole system. However, it may be useful for evaluators to replicate certain aspects of the system where they have come under criticism, or where certain aspects are seen as being particularly crucial or sensitive to typical operational variabilities. For instance, examples of this strategy (of replicating aspects of an assessment system) were utilized by the Human Resources Research Organization (HumRRO, 2009) in its evaluations of the California High School Exit Exam.

Concerns about this standard of (hypothetical) comparability become critical when aspects of an assessment system are not replicable. For example, when the data analysis for an assessment system is carried out using proprietary software, or other software that is not directly available and examinable by the evaluators, then a serious threat to transparency occurs. Another such problem would arise if an examination of the standard-setting exercise (described above in the section “Design of the Constructs in an Assessment System”) was found to reveal that the decision-making process was overly influenced by the presence of a particular individual in the standard-setting committee or if there was a wide variance in the judgments across the committee. In neither case does this issue mean that the assessment system is necessarily malfunctioning, but in both cases, a testing of the procedures would be called for: For example, it would be useful if standard-setting operations included general, blind replications for evaluating accuracy.

### CONCLUSION

Comparability is often seen as mainly a technical issue in discussions about assessment systems (e.g., how do scores on the ACT match to those on the SAT—an issue that is very prominent in university admissions policies). In the perspective adopted here, this is secondary—not that that reduces its importance—but the prior question must first be answered: Are the different parts of the system measuring the same or similar variables? We have ventured to respond to this question, addressing it from the perspectives of (1) the targeted subject-matter content, (2) the constructs in the tests, and (3) the stringency of the tests. This focus on the meaning of the tests is strategically important for users of the test results; there is little hope that establishing test-to-test concordances will be useful or valid if one does not have confidence that the tests address essentially the same underlying variables.

In thinking about comparability, we have also reviewed the issue of the relative comparability of results pertaining to different levels of aggregation—an issue that is often misunderstood, even among those with technical backgrounds. Another comparability issue that has not been thoroughly explored in the literature is the issue of transparency: What sorts of information should consumers have available to make decisions, and what level of technical documentation is needed to ensure that a system can be fully reviewed by expert evaluators?

One topic we have not covered in depth is the set of challenges in ensuring continuing comparability over time (i.e., historical years). This is an important issue for the practical usefulness of an assessment system. Some of these issues are parallel to issues we have raised above. Others are unique to this perspective. For example, item pools need to be maintained over time to ensure comparability of measurements. There needs to be refreshment. There may be reuse (which is also essential to check for comparability). What designs are good for this? How big must the pools be? It is also essential that there be release of real items (not rejected items), allowing use for teacher training, for instruction, and for general illustration of results. What are good strategies for this? How big does the release need to be? There should also be realistic practice tests. A further check could be the rescoring of the previous year's tests to check for rater drift.

Across the discussions in this chapter, there is a common thread of the need for strategic thinking about comparability. This must proceed from a clear statement of the expected uses of the results from a testing system, thus indicating where comparability is required and by whom (i.e., who the people are who will be making decisions). We see the need for clear statements about the underlying constructs as being the bedrock of such usages, and hence have given that issue (embedded as it is within decisions about content and stringency) priority here.

As we move forward, we expect to see new developments in testing practice that will raise further thorny issues. One such topic is the possibility of integrating "interim" or even classroom information into the assessment system (e.g., Wilson, 1994). We have only poked around the sides of this issue in this chapter. A comprehensive account of such an idea would need its own chapter. It raises issues not just of comparability of scale as we have been discussing above, but also issues of comparability of probity and security of the information, as well as comparability of the judges and judgments on which testing practice must be based. In the long run, we see a strong likelihood that systems will be built that incorporate such information sources, and recommend that there be a concerted effort to explore the many issues involved.

## REFERENCES

- Ackerman, T. A. (1988, April). *An exploration of differential item functioning from a multidimensional perspective*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.
- Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, 29, 67–91.
- Beaton, A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics*, 17(2), 191–204.
- Bejar, I. I., & Graf, E. A. (2010). Updating the duplex design for test-based accountability in the twenty-first century. *Measurement: Interdisciplinary Research and Practices*, 8(2–3), 110–129.

- Bennett, S. M., & Carlson, D. (1986). *A brief history of state testing policies in California*. Retrieved from <https://www.princeton.edu/~ota/disk2/1987/8724/872423.PDF>.
- Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. New York: David McKay.
- Bock, R. D., & Mislevy, R. J. (1981). An item response curve model for matrix-sampling data: The California grade-three assessment. *New Directions for Testing and Measurement*, 10, 65–90.
- Bock, R. D., Thissen, D., & Zimowski, M. F. (1997). IRT estimation of domain scores. *Journal of Educational Measurement*, 34(3), 197–211.
- Briggs, D. C. (2013). Measuring growth with vertical scales. *Journal of Educational Measurement*, 50(2), 204–226.
- CAASPP (California Assessment of Student Performance and Progress). (2015). *Assessment target reports frequently asked questions*. Retrieved from <http://www.caaspp.org/rsc/pdfs/CAASPP.target-report-FAQs.2016.pdf>.
- California Department of Education. (2019). *2018 California school dashboard technical guide final version: 2018–19 school year*. Sacramento, CA: Author. Available at <https://www.cde.ca.gov/dashboard>.
- CCSSI (Common Core State Standards Initiative). (2010). *Common Core State Standards for mathematics*. Retrieved from <http://www.corestandards.org/the-standards>.
- Cizek, G. J., & Bunch, M. B. (2006). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Los Angeles: Sage.
- Cogan, L. S., Schmidt, W. H., & Wiley, D. E. (2001). Who takes what math and in which track? Using TIMSS to characterize U.S. students' eighth-grade mathematics learning opportunities. *Educational Evaluation and Policy Analysis*, 23(4), 323–341.
- Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper & Row.
- Cronbach, L. J., Bradburn, N. M., & Horvitz, D. G. (1995). Sampling and statistical procedures used in the California Learning Assessment System. In L. J. Cronbach, *A valedictory: Reflections on 60 years in educational testing*. Washington, DC: National Academy Press.
- Cronbach, L. J., Gleser, G. C., Harinder, N., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: John Wiley & Sons.
- Darling-Hammond, L., & Pecheone, R. (2010, March). *Developing an internationally comparable balanced assessment system that supports high-quality learning*. National Conference on Next-Generation K–12 Assessment Systems. Retrieved from <https://www.ets.org/Media/Research/pdf/Darling-HammondPechoneSystemModel.pdf>.
- Dray, A. J., Brown, N. J. S., Diakow, R., Lee, Y., & Wilson, M. R. (2019). A construct modeling approach to the assessment of reading comprehension for adolescent readers. *Reading Psychology*, 40(2), 191–241. <http://doi.org/10.1080/02702711.2019.1614125>.
- Ebel, R. (1970). Behavioral objectives: A close look. *The Phi Delta Kappan*, 52(3), 171–173.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Gianopoulos, G. (2019). *From through-course summative to adaptive through-year models for large-scale assessment: A literature review*. Portland, OR: NWEA Research.
- Glaser, R. (1990). The reemergence of learning theory within instructional research. *American Psychologist*, 45, 29–39.
- Glass, G. (1978). Standards and criteria. *Journal of Educational Measurement*, 15(4), 237–261.
- Holzinger, K. J., & Swineford, R. (1939). The bifactor method. *Psychometrika*, 2, 41–54.
- Hoover, H. D. (2003). Some common misconceptions about tests and testing. *Educational Measurement: Issues and Practice*, 22(1), 5–14.
- Hoskens, M., & Wilson, M. (1999). *StandardMap* [Computer program]. University of California, Berkeley.
- HumRRO (Human Resources Research Organization). (2009). *Independent evaluation of the California High School Exit Examination (CAHSEE): 2009 evaluation report volume 1* (D. E. Becker, L. L. Wise, & C. Waters, Eds.). Retrieved from <https://edsources.org/wp-content/uploads/old/cahsee09evalrptv1.pdf>.
- ISO (International Organization for Standardization). (1994). *ISO 5725-1:1994(en): Accuracy (trueness and precision) of measurement methods and results—Part 1: General principles and definitions*. Retrieved from <https://www.iso.org/obp/ui/#iso:std:iso:5725:-1:en>.
- Kane, M. (1996). The precision of measurements. *Applied Measurement in Education*, 9(4), 355–379.

- Kohler, C., & Hartig, J. (2017). Practical significance of item misfit in educational assessments. *Applied Psychological Measurement*, 4(5), 388–400.
- Kolen, M. J., & Brennan, R. L. (2014). *Test equating: Methods and practices* (3rd ed.). New York: Springer-Verlag.
- Linn, R. L. (2005). *Fixing the NCLB accountability system* (CRESST Policy Brief 8). Retrieved from [http://www.cse.ucla.edu/products/policybriefs\\_set.htm](http://www.cse.ucla.edu/products/policybriefs_set.htm).
- Masters, G., & Forster, M. (1997). *Mapping literacy achievement: Results of the 1996 National School English Literacy Survey*. Hawthorn, Australia: ACER Press.
- Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects* (ETS Policy Information Center Rep. No. 90). Princeton, NJ: Educational Testing Service.
- NCES (National Center for Education Statistics). (2018). *What does the NAEP mathematics assessment measure?* Retrieved May 30, 2019, from <https://nces.ed.gov/nationsreportcard/mathematics>.
- NRC (National Research Council). (2001). *Knowing what students know: The science and design of educational assessment* (J. Pellegrino, N. Chudowsky, & R. Glaser, Eds.). Washington, DC: National Academy Press.
- NRC. (2013). *Next generation science standards: For states, by states*. Washington, DC: The National Academies Press.
- NRC. (2015). *Guide to implementing the Next Generation Science Standards*. Washington, DC: The National Academies Press.
- PARCC (Partnership for Assessment of Readiness for College and Careers). (n.d.). *Performance levels: Knowledge, skills, and practices*. Retrieved May 30, 2019, from <https://parcc-assessment.org/performance-levels>.
- Preston, J., & Moore, J. E. (2010, March). *An introduction to through-course assessment*. Raleigh, NC: North Carolina Department of Public Instruction. Retrieved from <http://www.dpi.state.nc.us/docs/intern-research/reports/through-course.pdf>.
- Resnick, L. B., & Berger, L. (2010, March). *An American examination system*. National Conference on Next-Generation K–12 Assessment Systems. Retrieved from <http://www.k12center.org/rsc/pdf/ResnickBergerSystemModel.pdf>.
- Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments* (pp. 37–76). Boston, MA: Kluwer Academic Publishers.
- Roberts, L., & Sipusic, M. (1999). *Moderation in all things: A class act* [Film]. Available from the Berkeley Evaluation and Assessment Center, Graduate School of Education, University of California, Berkeley, Berkeley, CA 94720-1670.
- Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, 22, 53–61.
- Schmidt, W. H., McKnight, C., Houang, R. T., Wang, H. A., Wiley, D. E., Cogan, L. S., & Wolfe, R. G. (2001). *Why schools matter: A cross-national comparison of curriculum and learning*. San Francisco, CA: Jossey-Bass.
- Schmidt, W. H., McKnight, C., & Raizen, S. (1997). *A splintered vision: An investigation of U.S. science and mathematics education*. Dordrecht, the Netherlands: Kluwer.
- Shepard, L. (1980). Standard setting issues and methods. *Applied Psychological Measurement*, 4(4), 447–467.
- Suter, L. E. (2017). How international studies contributed to educational theory and methods through measurement of opportunity to learn mathematics. *Research in Comparative and International Education*, 12(2), 174–197.
- Tyler, R. W. (1934). A generalized technique for constructing achievement tests. In *Constructing achievement tests*. Columbus, OH: Bureau of Educational Research.
- Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, 24, 185–202.
- Walker, D. A. (1962). An analysis of the reactions of Scottish teachers and pupils to items in the geography, mathematics and science tests. In A. F. Foshay, R. L. Thorndike, F. Hotyat, D. A. Pidgeon, & D. A. Walker (Eds.), *Educational achievements of thirteen-year-olds in twelve countries*. Hamburg, Germany: UNESCO Institute for Education.
- Wang, M. (1986, April). *Fitting a unidimensional model to multidimensional item response data*. Paper presented at the Conference of the Office of Naval Research Contractors, Gatlinburg, TN.



- Wilson, M. (1994). Community of judgement: A teacher-centered approach to educational accountability. In *Issues in educational accountability*. Washington, DC: Office of Technology Assessment, U.S. Congress.
- Wilson, M. (1997). The California comparability study. In *Proceedings: Comparability symposium*. Downey, CA: Los Angeles County Office of Education.
- Wilson, M. (2004). A perspective on current trends in assessment and accountability: Degrees of coherence. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability*. 103rd yearbook of the National Society for the Study of Education, Part II (pp. 272–283). Chicago, IL: University of Chicago Press.
- Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.
- Wilson, M. (2009, December). *Assessment for learning AND for accountability*. Keynote presentation at the Exploratory Seminar: Next Generation K-12 Assessment Systems. Educational Testing Service, Princeton, NJ.
- Wilson, M. (2018). Making measurement important for education: The crucial role of classroom assessment. *Educational Measurement: Issues and Practice*, 37(1), 5–20.
- Wilson, M., & Draney, K. (2002). A technique for setting standards and maintaining them over time. In S. Nishisato, Y. Baba, H. Bozdogan, & K. Kanefugi (Eds.), *Measurement and multivariate analysis: Proceedings of the International Conference on Measurement and Multivariate Analysis, Banff, Canada, May 12–14, 2000* (pp. 325–332). Tokyo: Springer-Verlag.
- Wilson, M., & Gochyyev, P. (2020). Having your cake and eating it too: Multiple dimensions and a composite. *Measurement*, 151, 107247. <https://doi.org/10.1016/j.measurement.2019.107427>.
- Wise, L. L. (2011). *Picking up the pieces: Aggregating results from through-course assessments*. Center for K–12 Assessment & Performance Management at ETS. Retrieved from [https://www.ets.org/Media/Research/pdf/TCSA\\_Symposium\\_Final\\_Paper\\_Wise.pdf](https://www.ets.org/Media/Research/pdf/TCSA_Symposium_Final_Paper_Wise.pdf).
- Wolfe, R. G. (2000). *Opportunity to learn: Little-o and Big-O*. Presentation at the CCSSO Large-Scale Assessment Conference, Snowbird, UT.
- Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 14, 97–116.
- Zwick, R., & Mislevy, R. J. (2011). *Scaling and linking through-course summative assessments*. Center for K–12 Assessment & Performance Management at ETS. Retrieved from [https://www.ets.org/Media/Research/pdf/TCSA\\_Symposium\\_Final\\_Paper\\_Zwick\\_Mislevy.pdf](https://www.ets.org/Media/Research/pdf/TCSA_Symposium_Final_Paper_Zwick_Mislevy.pdf).





# Comparability Across Different Assessment Systems

Marianne Perie, *Measurement in Practice, LLC*

## CONTENTS

INTRODUCTION .....	123
REQUIREMENTS FOR COMPARABILITY .....	124
Purpose. ....	125
Content. ....	125
Administration Conditions .....	127
Psychometric Characteristics .....	128
COMMON PRACTICE .....	128
LEVELS OF COMPARABILITY .....	130
EXAMPLES .....	130
Interim to Summative Assessments. ....	130
State High School Assessment to College-Readiness Exam. ....	133
Cross-State Comparisons. ....	136
National Assessments Compared to International Benchmarks .....	142
Different International Assessments Compared .....	143
CONCLUSION .....	146
REFERENCES .....	147

## INTRODUCTION

With the production of the Common Core State Standards (CCSS), many state and national policy makers indicated a desire to compare performance across states and jurisdictions. For example, could the performance in Florida be compared to that in Illinois or Texas, and could performance in Miami be compared to that in Chicago or Dallas? The National Assessment of Educational Progress (NAEP) provides comparable scores at the state level and for some urban districts; however, state leaders have indicated the desire for comparability that goes beyond the level that is currently available. On the one hand, they wanted to compare their school or district to top achieving countries in the world, and on the other, they wanted to be able to understand equivalent scores from students coming into their school from another state. In an early meeting,

prior to the development of the multistate testing consortia funded by the federal government under the Race to the Top funding statute, state leaders verbally indicated the following reasons for wanting comparable scores:

- To understand the performance of a student transferring from another state;
- To compare schools across states;
- To know that the term “college ready” has the same meaning from one state to another;
- To compare proficiency across schools, districts, and states; and
- To compare performance in a district to performance in another country.

As policy has moved the focus to college readiness, there is also a desire to compare state assessments to the tests traditionally used for college admissions: the ACT and the SAT. The consortia leaders wanted to replace those tests with their own high school equivalents, and states wanted to be able to predict ACT and SAT scores using their state assessments.

Finally, to achieve all of these outcomes, instruction needed to be aligned with the summative goals. Districts often purchase interim assessments that are purported to align with their state standards to predict success on the summative assessment and improve outcomes.

All of these goals require a degree of comparability among each of these assessments depending on how they are used to meet each goal. Many researchers have struggled with the question of how to compare scores across tests. For example, Braun and Qian (2007); Mislevy (1992); and NRC (1999a, 1999b) have produced important publications documenting the challenges to and potential approaches for comparing scores across states using NAEP, state tests, and other measures. The central challenges that Mislevy outlined in *Linking Educational Assessments* remain:

- “discerning the relationships among the evidence the assessments provide about conjectures of interest and
- determining how to interpret this evidence correctly” (p. 21).

In this chapter, we examine how different assessment systems can be designed to answer questions about comparability of students, schools, districts, and states. Specifically, the focus is on required elements of the assessments for comparability, understanding the comparability of scores at different levels of aggregation, and psychometric constraints in making the desired inferences about students and schools across states and countries.

## REQUIREMENTS FOR COMPARABILITY

Returning to the definition of comparability used in this volume, scores are comparable when they can be validly related even when they come from measurements taken at different times, in different places, or using variations in assessments and assessment procedures. Ideally, users could be assured that students with the same score on the same scale possessed the same level of proficiency with respect to the domain of knowledge and skills that a test was intended to measure.

Although the requirements for comparability differ depending on the comparisons made, there are some basic principles that should be followed. Variations in the degree of similarity of these principles affect the degree of comparability that can be assumed: first is the stated purpose of the tests, which affects the test takers' motivations and often differs across tests even within the same system; second, the similarity of the content of the assessments influences what can be said about the scores; third, administration conditions can affect the degree of comparability; and fourth, the psychometric properties of the assessments may change the interpretability of any comparison.

### **Purpose**

Depending on the intended use of the test, a student may spend more or less time preparing for it and may give varying levels of effort during the test-taking session. There is a large body of research documenting the correlation between motivation and achievement (e.g., Finn, 2015; Pintrich, 1989; Pintrich & DeGroot, 1990). Motivation can affect the score because unmotivated students may not give their full effort, resulting in a score that underestimates their actual knowledge and skills (Mislevy, 1992). Even when there are no stakes for the students, such as state assessments under the No Child Left Behind (NCLB) era, the results may affect teachers, which could lead to increased test preparation activities.

Purpose could also affect performance, for example, when the difference is between a benchmark test used to inform the teacher on where to start teaching the student and a summative assessment that "grades" students or their teachers. Putting motivation aside, the interim assessment is likely to be a part of classroom instruction and to have less fanfare associated with it across the school. Because the results can be shared immediately with the student, the test could be perceived differently, and the student may put forth a greater effort than with many state assessments.

Comparing results from a test with no stakes, and a likely proportion of unmotivated test takers, to one with high stakes and highly motivated test takers can result in a large degree of error. For example, predicting a college-readiness score on a test like the ACT from a grade 10 state assessment with no student-level consequences could result in underpredicting performance by several points.

### **Content**

Typically, content similarities are shown through alignment studies that compare the specific content and depth of knowledge assessed by the items on one assessment to that of another (see, e.g., Webb, 1997, 2007). However, traditional alignment studies simply document the percentage of items that can be mapped directly to a standard, the percentage of standards that have items measuring them, the range of depth of knowledge of the items across the standards, and the balance of representation of items across the standard. Two tests can receive similar alignment ratings to the same set of standards and still measure the subject area quite differently. Conversely, two tests could be very similar in what they measure and how they measure it and not receive the same alignment score. Alignment is too superficial as currently defined to be a sole requirement for content comparability.

Although strict content alignment is not required for comparability, the tests should measure the same construct at the same level. Tests that differ in terms of what is assessed or even the distribution of emphasis on the knowledge and skills assessed can affect the comparability of the two scores. When the two consortia met to discuss comparing scores across the two sets of tests, some argued that building the test to the same content standards was a necessary but insufficient condition for comparability. Additionally, the tests would need to use the same blueprints that emphasized the same content areas as well as the same performance-level descriptions (e.g., Marion & Perie, 2011). Even if all that was desired was to claim that the percentage of students who reached proficiency could be compared across all states using one of the consortia assessments, the same definition of proficiency would need to be adopted for each. Using the same method to set the cut scores for proficiency and incorporate the same external benchmarks would also strengthen comparability claims. Ultimately, none of this was done, and scores from the two consortia are not considered interchangeable.

Going further than test specifications or blueprints and performance expectations, item types can also influence the score interpretation. Two tests purporting to measure the same construct but using different item types may be measuring similar content at different levels of rigor. Consider a math test that is all multiple choice versus one that contains additional items asking students to show their work and explain their reasoning. Although both of those constructs could be measured with selected-response items, asking a student to generate a response taps into a different psychological construct that may affect the comparability of the results while providing different insights into the student's learning.

Going even further, if two tests both include open-ended items but are scored with rubrics that emphasize different aspects of the construct, those results could also be limited in comparability. Consider, for example, a rubric that focuses on the quality of evidence used to defend an argument compared to one that focuses on the organization of the paragraph. And, even if the rubrics are similar, differences in scoring processes could also affect comparability, as discussed in the next section.

Finally, even when tests are designed based on the same specifications and using the same performance-level descriptions, differences in alignment of instruction to those standards can influence statistical linking of assessments. For example, Reardon, Kalogrides, and Ho (2018) showed that the linkage between state tests and NAEP showed higher-than-expected scores on the state tests than indicated by linked NAEP scores for districts participating in the Trial Urban District Assessment. One explanation they provide is that the district instruction may be more closely aligned with the state test than that of other districts. When the linkage between the state tests and NAEP is done at the state level, district means predicted from one test to the other have differing degrees of accuracy because curriculum and instruction vary by district.

One alignment report by Forte (2017) discusses the level of alignment needed for comparability, as she describes an approach that starts with the claims made about the resulting scores and traces them through the content standards, blueprints, specifications, and performance-level descriptors. Her theory is that all pieces of an assessment must be aligned to claims made about the student, classroom, school, or state in the final reports.

### Administration Conditions

The degree of comparability between two scores also depends on the conditions under which the tests were administered. An issue that states struggled with within each consortium was the degree of flexibility that should be allowed for administration. Take, for instance, the testing window. Some states traditionally set aside one week in the spring for testing, and all assessments are completed that week under strict schedules. Other states have much longer windows, and districts and schools have the flexibility to schedule the assessments within that window. Even with the same length of window, some testing programs assess earlier in the spring than others. Allowing for different assessment windows could have the effect of providing fewer or more instructional hours prior to the assessment, which would negatively affect the comparability of the assessment scores.

A second type of administration condition related to time is the speededness of the test. Although the research is mixed about the direction and significance of this effect, whether the test is timed can affect performance (Haniff, 2012). Therefore, equating error is introduced if we try to link scores taken in a timed test condition to those taken under untimed conditions.

Additionally, the same assessment could be given at the same time through multiple platforms. Some districts may provide the assessment using paper and pencil, others on a personal computer, and others on a tablet; still others may use a combination of platforms. Multiple studies over the past few years have focused on the comparability of paper and pencil to computer-based assessment and on device comparability within computer-based assessment (see, e.g., DePascale, Dadey, & Lyons, 2016; Kingston, 2009; Way, Davis, Keng, & Strain-Seymour, 2016). Differences have been found between paper and pencil and computer assessments as test designers take advantage of technology in ways that can be difficult to translate to paper and pencil. Among technology devices, however, few differences have been found. As long as the students are familiar with the device on which they take the test, the results are assumed to be comparable across devices.

A fourth difference in administration conditions that can affect comparability is the accommodations allowed. Although there is now general consensus on most accommodations, there continue to be different policies on when a read-aloud accommodation is used and when calculators may be used by students with disabilities.<sup>1</sup> There is almost universal agreement that students may have instructions read aloud, regardless of whether they have a disability. Likewise, there is general agreement that reading aloud a math item to those who need that form of communication does not alter the construct being assessed. However, there exist greater differences in opinion about when and how the read-aloud accommodation should be permitted in an English language arts (ELA) assessment (Rogers, Lazarus, & Thurlow, 2014). Some policy makers do not allow it until students have reached a certain grade level so that decoding can be measured in the lower grades. Others argue that reading aloud any part of a reading test at any grade changes the construct assessed and should not be allowed. Still others

---

<sup>1</sup> For further discussion on accommodations, please refer to Chapter 6, Comparability When Assessing English Learner Students, and Chapter 7, Comparability When Assessing Individuals with Disabilities, in this volume.



believe that we need to assess a student in any way that elicits information from them and thus they allow the read-aloud accommodation at all grades. These differences of opinion manifest themselves in different accommodations policies found across states and other jurisdictions. These differences mean that the scores across jurisdictions with different policies would not be comparable for students needing the accommodation.

### Psychometric Characteristics

Other factors of the assessment can also affect the degree of comparability of the scores. Both tests being compared should have similar, high reliabilities in order to make interpretable comparisons. A low reliability on either test would increase the error in linking them and increase the confidence interval around the linked score.

Likewise, the model used to scale the assessments should be the same. If one test uses a one-parameter model to scale the tests and the other uses a three-parameter model, the linkage will have a greater degree of error. When equating two forms of a test, items are calibrated together using the same model. But, often, in trying to project the score from one test onto another, recalibration is not an option. For example, when trying to link a state end-of-course math assessment to the SAT, one of the factors resulting in a high degree of error was that the state test used an item response theory model to create its scale while the College Board uses a classical approach that norms the results (Roeber et al., 2018).

Finally, as discussed in the section on content, different item types can affect comparability. Moreover, within constructed-response item types, different scoring rules can also affect the results, usually by increasing or decreasing the reliability of the overall assessment. Clearly, the rubrics themselves can affect comparability. If different content expectations are emphasized in the rubric, it can reduce the clarity of score interpretation when the two scores are compared. Furthermore, if one program uses a rigorous scorer training protocol with frequent validity checks and read behinds, while the other relies more on remote training with few checks, the resulting score discrepancies can affect the comparability of the results. Likewise, the density of items around specific points in the scale affects reliability, so tests designed to spread items across the scale will have different precision at specific points than will tests designed to maximize information at a specific cut score of the scale.

Ultimately, the comparison between tests could be made at the score level or at some benchmark. For example, under NCLB, states wanted to compare the percentage of students reaching “proficient.” Under the revised act, called the Every Student Succeeds Act (ESSA), the focus is more on comparing the percentage of students who are “college ready.” And while one could argue that “proficiency” means “ready for college,” the same terms may not have the same definitions across different states. And without the same definitions, the percentages are not comparable.

### COMMON PRACTICE

Regardless of the principles of comparability, statistically, most tests can be linked. Multiple researchers have linked scores from state assessments to NAEP (Bandiera de Mello, Blankenship, & McLaughlin, 2009; Braun & Qian, 2007), from NAEP to the

Programme for International Student Assessment (PISA) (Stephens & Coleman, 2007) and Trends in International Mathematics and Science Study (TIMSS) (Jia et al., 2014), from one state to another (Bandeira de Mello, Rahman, & Park, 2018), from interim assessments to summative assessments (Reardon, Kalogrides, & Ho, 2018), and from state or local assessments to the ACT and the SAT (Roeber et al., 2018). Test score equating or linking, the more general term, is the most common way we address comparability goals in our current testing context.

The goal of equating is to disentangle differences (across different forms or tests) in item or form difficulty from changes in actual student achievement. A common current example is ensuring that the scores on the state's fifth-grade mathematics test in 2019 can validly be placed on the same scale as the 2018 scores. In this example different students (the fifth graders in 2018 and 2019) have completed tests containing different sets of items, except for a subset of items that were administered in both 2018 and 2019. It is this subset of items—assuming many conditions are met—that allows us to disentangle the changes in student achievement from the changes in the difficulty of the other (nonlinking) items on the test. The challenge is to ensure that the assumptions are actually met.

Although equating is the strongest form of linking, it can only be conducted when the two tests were designed from the same test blueprint to measure the same construct(s). Holland (2007) describes the purpose of equating as making it possible to use scores interchangeably, which can result when the tests measure the same construct with the same intended difficulty and reliability. The most common example is to use two or more forms of the same test. This is not the level of comparability of interest in this chapter, so equating is not discussed here.

The two most common forms of linking when comparing one test score to another different score are calibration and projection. Calibration is used when the tests were not designed from the same test blueprint, but both have been constructed to provide evidence about the same type of achievement (e.g., the same construct). “Unlike equating, which matches tests to one another directly, calibration relates the results of different assessments to a common frame of reference, and thus to one another only indirectly” (Mislevy, 1992, p. 24). Calibration is described as a type of scale aligning with the purpose of “transforming scores from two different tests onto a common scale” (Holland, 2007, p. 12). Projection is used to make statements like “a student who scores X on Test A would have a 75% probability of scoring between Y and Z on Test B.” It has a looser set of requirements for the comparability of the two tests, but, as described throughout this chapter, when assessments are constructed around different types of tasks, administered under different conditions, or used for purposes that bear different implications for test takers' affect and motivation, then mechanically applying linking or aligning formulas can prove seriously misleading (Mislevy, 1992).

More recently, the term “concordance” has been used to refer to linking scores on assessments that measure similar (but not identical) constructs and in which scores on any of the linked measures are used to make a particular decision (Kolen, 2004). Subsumed in this definition is the assumption that scores are highly correlated and test takers are similar. For example, ACT and the College Board design studies that result in concordance tables for the ACT and the SAT, which allow one to determine the equivalent score on the test not taken based on the score of the test taken (College

Board & ACT, 2018). Dorans (2004) recommends using regression methods to link scores on measures that cannot be related using concordance procedures. Additional detail on linking can be found in Chapter 2, Comparability of Individual Students' Scores on the "Same Test."

### LEVELS OF COMPARABILITY

Thus, the question is no longer how to link one test to another but how to interpret the results and determine if they are truly comparable. Examining the error associated with the linkage will tell us the precision with which we can estimate what score a student would have likely received if they had taken the other test. However, policy makers may be more interested in how scores from one group of students compare to scores from another group. For instance, can we compare Algebra I performance in Los Angeles to that in Chicago when the two districts take two different state end-of-course assessments? Moreover, the comparison might not be made in terms of average scale score but in terms of the percentage of students who "pass" or reach a specific standard.

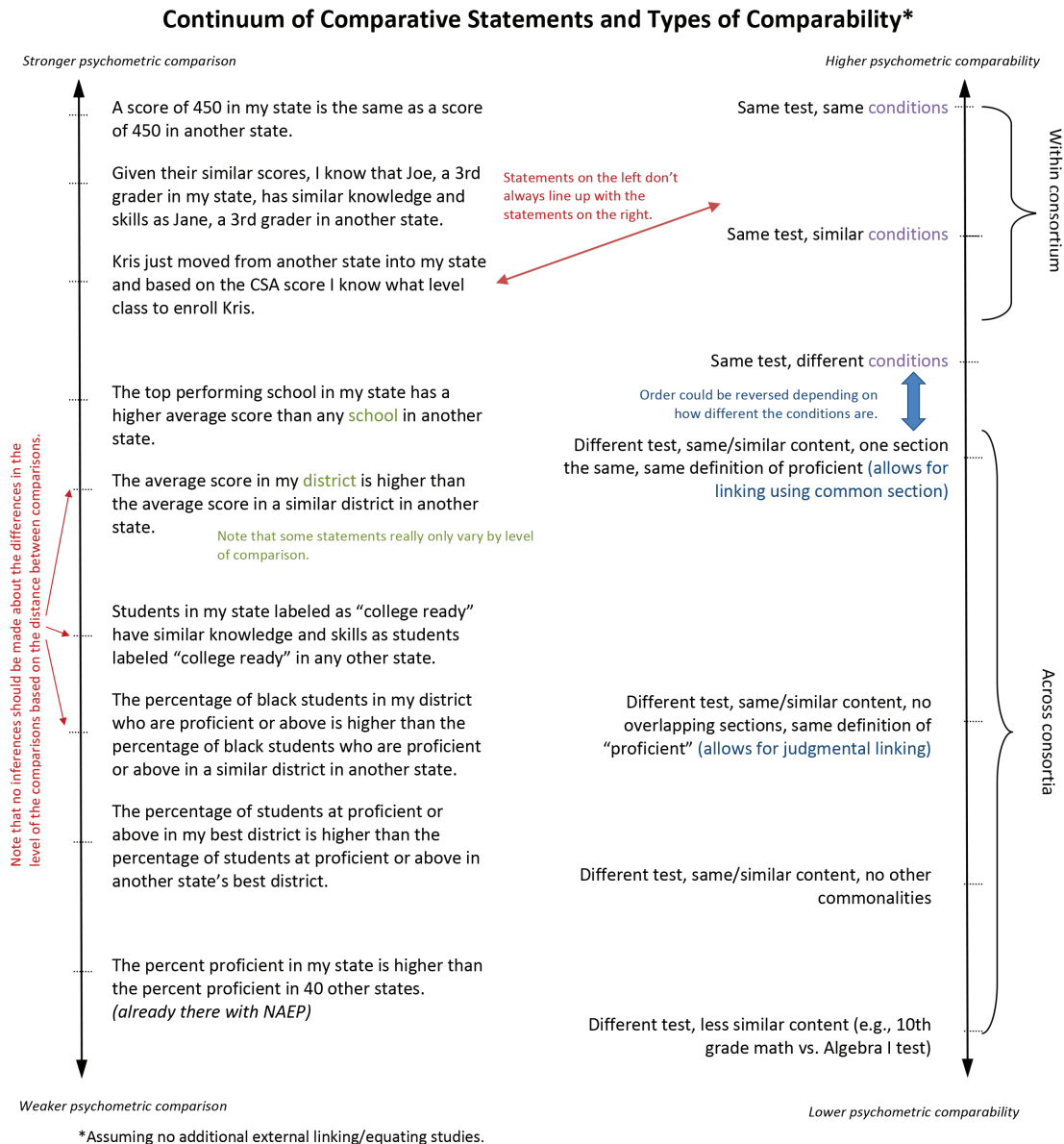
Figure 5-1 graphically presents a selection of statements that one might want to make about linked scores (Marion & Perie, 2011). As can be seen from the figure, the strictest student-level comparability requires the same test to be administered under the same conditions. The authors of the figure acknowledge that it likely oversimplifies the ordering of the statements and assessment conditions and that the order could change slightly as they attempt to display several factors on a single line. The figure provides a good general overview of the trade-offs between comparability statements and design and administration conditions but should not be viewed as a menu. In the next section, specific examples of different types of linkage are given along with levels of comparability attained.

### EXAMPLES

Next, we walk through examples of the ways two different tests are linked to create comparable scores and examine how comparable they truly are. The important piece is the claim being made after scores are linked. The claim that scores are interchangeable requires much more rigorous conditions of comparability than the claim that the rank order of schools would be the same on two assessments or that the percentage of students reaching a benchmark would be comparable.

#### Interim to Summative Assessments

The creation of consortia also led to the ability of states to purchase balanced assessment systems. Smarter Balanced continues to provide summative assessments, interim assessments, and formative resources. Chapter 4, Comparability Within a Single Assessment System, discusses comparability within such systems. However, only 14 states and 2 territories remain in either consortia at the time of this writing. Most states develop their own summative assessments through a contractual process with a test development contractor. The development or purchase of interim assessments and formative tools are then left to the districts. This practice can lead to a large amount of local



**FIGURE 5-1** Continuum of comparative statements and level of comparability.  
 SOURCE: Marion & Perie (2011).

variation in the comparability of scores between interim and summative assessments. Some states (e.g., Arizona, Florida, and Louisiana) review and approve interim assessments as being sufficiently aligned with the state standards and of sufficient technical rigor to produce results that should be in line with information from the summative assessment. Other states leave to their districts the task of reviewing the technical quality and comparability of the interim assessments.

Many interim assessment companies tout their products as useful in improving scores on summative assessments (Perie, Marion, & Gong, 2009). In order to make claims about how growth on the interim assessment leads to higher summative assessments, they link the two tests. Typically, there are students taking both tests, and a regression equation is set up to predict the score on the summative assessment from the interim assessment. Depending on how close in time the interim assessment is administered to the summative assessment, the prediction can be fairly accurate (see, e.g., Immekus & Atitya, 2016). However, according to Li, Marion, Perie, and Gong (2010), there are several other conditions that should be met to enhance the comparability of the assessments.

### *Purpose*

The interim assessment typically serves a much different purpose than a state summative assessment. An interim assessment is intended to provide instructional feedback to teachers that can lead to corrective action and/or be used to measure growth in understanding over smaller bits of time (Perie et al., 2009). Students are more likely to be engaged because they receive immediate feedback, and the results will affect their learning opportunities. The state summative assessment typically takes months to produce student scores and they may affect school and teacher accountability. However, interim assessments could also be used by policy makers to ensure teachers follow a similar scope and sequence of instruction or to predict performance on the summative assessment, which could yield less desirable results.

### *Content*

Chapter 4, *Comparability Within a Single Assessment System*, discussed the situation where an interim assessment is built as part of a balanced assessment system. In that case, the standards and item specifications should be the same. However, in many districts, curriculum directors buy an off-the-shelf product that appears to align with their standards and curricular goals.

Many interim assessment companies have spent the past several years aligning their content to the CCSS. During that same time, many states have revised their standards and thus they can no longer be called Common Core State Standards. Depending on the degree of changes made by the state, interim assessments will be more or less aligned to the state standards. Some companies allow districts to select items (or standards) from an item bank to build assessments, a practice that should lead to better alignment. As discussed by Perie et al. (2009), interim assessments can be built as miniature summative assessments, covering all standards, or as a section of the content that reflects a small number of standards that were expected to be taught by that point in the year. The

alignment between the expected scope and sequence of the interim assessment and the actual scope and sequence of the instruction will affect the interpretability of the results.

Finally, differences in item type can affect comparability. Interim assessments that serve the primary purpose of diagnosing a student's understanding may have more open-ended and probing questions. Interim assessments can consist solely of performance tasks, but those are rarely assessed on a large-scale test because of the cost of scoring them. Moreover, to serve a strong predictive function, Perie et al. (2009) argue that the item types on an interim assessment should match the item types on the summative assessment. So, those tests that provide rich instructional feedback are less comparable to the summative assessments and provide less reliable predictive information.

### *Administration Conditions*

If an interim assessment is built or purchased as part of a balanced assessment system, it is likely to have the same administration conditions. Typically, in those cases, the assessments are given on the same computer platform and students receive the same accessibility tools and accommodations every time they access the system. However, when an interim assessment is purchased separately, it is typically given on a different platform and may not include all of the same tools or accommodations. If the summative assessment has more rigorous administration conditions, the interim assessment may overpredict performance on the summative assessment. Conversely, if the interim assessment system does not include all of the necessary accommodations for a student, it may underpredict performance on the summative assessment for that student.

### *Psychometric Characteristics*

Typically, interim assessments have lower reliability than summative assessments because they tend to be shorter. However, the reliability still tends to be well in the acceptable range. Scaling models will often differ because the tests are developed and analyzed by different vendors. However, as long as the standard error associated with the prediction is reported clearly, the results should still be interpretable. As described by Immekus and Atitya (2016), the total score of an interim assessment is typically the best predictor of performance on the summative assessment. Subscores provide little additional value to the prediction equation or comparability.

## **State High School Assessment to College-Readiness Exam**

Currently, several states are replacing their high school assessments with either the ACT or the SAT and using that test for the dual purpose of high school accountability and a measure of preparedness for college coursework. Students can use the scores they receive on the state-administered test for admission into college. Other states are funding one administration of either the ACT or the SAT for every high school student but not using it for school accountability. But, for the states that are allowing the ACT or the SAT to be used as an alternative to their high school assessment, claims they make about the comparability of the two should be examined.

As with interim assessments, some states link their state assessment through a common student approach to make claims about expected performance on the ACT or



the SAT. For example, Kansas provides information on where the equivalent of an ACT benchmark falls on its statewide summative assessments in ELA and math at grades 8 and 10. It also created the linkage shown in Table 5-1 displaying likely ACT scores for student scores at each performance level of the Kansas Assessment Program.

**TABLE 5-1** Projected ACT Scores for Grade 10 KAP Performance Levels

English Language Arts		
KAP	ACT Reading	ACT English
Level 1: 220–268	1–17	1–16
Level 2: 269–299	18–23	16–22
Level 3: 300–333	23–29	22–28
Level 4: 334–380	29–36	28–36
Mathematics		
KAP	ACT	
Level 1: 220–274	1–19	
Level 2: 275–299	19–22	
Level 3: 300–332	23–27	
Level 4: 333–380	28–35	

NOTE: KAP = Kansas Assessment Program.

SOURCE: <https://ksassessments.org/act>.

As with interim assessments, the college-readiness assessments were written to different standards and specifications. However, with a common-population linking design, regression equations can be used to predict performance on one test given performance on the other. More information about the error involved should be provided to help interpret the scores, but the test scores are not intended to be used interchangeably.

A different situation arose in Florida, which had legislation that required the state to analyze the possibility of replacing the Florida State Assessment (FSA) high school ELA test and the Algebra end-of-course assessment with either the ACT or the SAT. The decision would be made at the school level, but then Florida would compare schools based on scores placed back on the FSA scale. In this case, the scores from the FSA, ACT, and SAT would have been considered interchangeable, which requires a high level of comparability. A report commissioned by a group of researchers showed that the criteria needed for this level of comparability could not be met (Roeber et al., 2018). Although the SAT was fairly well aligned to the state content standards, the ACT would need to be supplemented to measure the same detail as measured by the FSA. More importantly, statistical linking showed a large degree of error in trying to predict scores from the FSA to the ACT and the SAT or vice versa. Because the results would be used for school accountability purposes, decision accuracy was calculated using the ACT and the SAT and assuming the FSA decision was “correct.” Results of this classification consistency analysis indicate that many students would be placed at different performance levels on the three tests, some by as much as four out of the five performance levels. Particularly concerning was that the direction of the error varied depending on the ability level of the students. Larger schools with a greater number

of lower-performing students have an advantage in using the alternate tests (the ACT and the SAT). Schools with a higher-performing population fared better when graded using the FSA. Florida ultimately withdrew legislation to allow the three assessment programs to operate as if the scores were interchangeable.

The most common approach currently for states is to allow for the “local option” clause in the ESSA. That is, a district can choose to give a “nationally-recognized college-ready assessment” in place of the state high school assessment if it meets peer-review requirements, including content alignment. At least a dozen states are using either the ACT or the SAT as the high school assessment used for accountability (Gewertz, 2019). A couple of states are allowing districts to choose the other assessment as the local alternative, if they desire. For instance, Oklahoma elected to have the SAT become its new state assessment in high school, replacing its end-of-course exams. But it also allows districts to choose to use the ACT instead and rely on the College Board/ACT concordance tables to place all the scores on the same scale. At the time of this writing, no state had both assessments pass peer review. A big issue is the comparability claim and conditions that must be met for it to be true.

### *Purpose*

Both the ACT and the SAT serve the same purpose of informing college admissions offices of a student’s level of knowledge, skill, and reasoning ability. Interestingly, though, they both now serve the additional purpose of school accountability at the high school level. A state summative assessment serves the latter purpose, but because there are no other stakes attached, students may be less motivated to do their best on the state assessment. There should be no difference in motivation or effort on the SAT compared to the ACT.

### *Content*

Beginning in March 2016, the College Board administered a new version of the SAT that was revised to better align with the CCSS. The ACT does not align to any particular standards, and its focus is on a framework rather than alignment to standards. Alignment studies have shown differences in the content covered by the two assessments, particularly in mathematics. The SAT includes items on linear equations, systems, problem solving, data analysis, complex equations, geometry, and some trigonometry. The ACT assesses pre-algebra, elementary algebra, intermediate algebra, plane geometry, and coordinate geometry.

The Florida study described earlier (Roeber et al., 2018) showed stronger alignment between the SAT and Florida’s state standards than between the ACT and the state standards, indicating a mismatch in content coverage between the SAT and the ACT. Achieve (2018) conducted an alignment study of the ACT to the CCSS and found a weak match between the two. Earlier, the Delaware and Maine State Departments of Education commissioned a study by the Human Resources Research Organization on the alignment of the SAT with the CCSS (Nemeth, Michaels, Wiley, & Chen, 2016). They found fairly strong alignment for ELA and slightly lower alignment for math, although it was concluded that the SAT met the minimum requirements for an aligned high-quality assessment.

These and other alignment studies often conclude with a recommendation that a state augment the college-readiness assessment with a set of items that cover the standards not assessed by the ACT or the SAT. While this approach can close gaps in content alignment, it becomes trickier to ensure these supplemental items are scaled appropriately and lead to valid score interpretations.

### *Administration Conditions*

The ACT and the SAT are administered on different platforms. Although they are now both offered as computer-based tests, different software is used to administer them. The majority of accessibility and accommodation options are the same. The primary difference is in reading directions aloud for English learner students. However, policies are continually reviewed and updated by both companies, so information here may no longer be accurate. The bigger difference is often between state administration policies and the ACT and the SAT policies. Both the ACT and the SAT are timed tests while the majority of state tests are not, and some states allow for different accommodations than others.

### *Psychometric Characteristics*

The comparability report commissioned by the Florida State Department of Education included a table comparing the psychometric properties of an administered form of the ACT, the SAT, and the equivalent FSA. The table is reproduced as Table 5-2 (Roeber et al., 2018, p. 78).

As seen in Table 5-2, there are more differences with the FSA than between the ACT and the SAT. They have similar reliabilities and mean item difficulties. The SAT has greater variability in item difficulty and includes grid-in items in the mathematics test. The ACT is a slightly longer assessment.

Even though the College Board and the ACT work diligently, using a common student approach, to link the scores of the two assessments, the level of comparability is not rigorous enough to assume the scores are as interchangeable as they appear in the concordance tables. The two tests do measure different content, particularly in mathematics, using different item types. And, certainly, they are quite different from typical state summative assessments.

### **Cross-State Comparisons**

A desire of many state commissioners is to compare performance in their state to others. Some describe practical reasons, such as being able to place the student's score from a different state assessment onto their scale to help with placement. Others simply want to raise their relative ranking. The consortia were born of the desire to have scores that can be transported as well as compared across states. Although one consortium, the Partnership for Assessment of Readiness for College and Careers (PARCC), has lost the vast majority of its state participants as of this writing, the other, Smarter Balanced, maintains sufficient numbers of states that comparisons can be made.

**TABLE 5-2** A Comparison of Form Characteristics of Florida State Assessment, ACT, and SAT

	ELA			Math		
Criterion	FSA: ELA 10	ACT	SAT	FL EOC: ALG I	ACT	SAT
Form reliability	0.91	0.89	0.89	0.92	0.91	0.90
Form length	53 items + writing prompt	115 items + writing prompt	96 items + writing prompt	58 items	60 items	58 items
Distribution of item types	58% MC; 23% Editing text choice; remaining is multiselect, hot text, and evidence-based Selected Response	MC + essay	MC + essay	Vast majority of item types are MC and SCR (SCR = grid-in or equation editor). Other = table and matching	MC	MC + grid-in
Item difficulty <sup>a</sup>						
Mean	0.65	0.58	0.58	0.21	0.58	0.58
Min	0.12	0.20	0.03	0.00	0.20	0.03
Max	0.92	0.89	0.98	0.75	0.89	0.98

NOTE: FL EOC = Florida end-of-course exams; MC = multiple-choice item, with four options and one correct answer; SCR = short constructed-response item.

<sup>a</sup> Item difficulty is shown as the percentage of students answering an item correctly. The minimum and maximum show the percentage of students answering the hardest and easiest item on a form correctly. The mean gives an indication of the overall difficulty of the form by summarizing the percentage of items answered correctly.

### *State to State on Consortia Assessments*

On the surface, it appears that a consortium of states giving the same assessment should have full comparability of the scores. That was certainly the intent of forming the consortia. Indeed, the consortia tests generally had the same purpose, content, and psychometric characteristics. However, administration conditions were not always the same.

Working within either consortium during development raised many issues of comparability. In order to have interchangeable scores across states, many administration decisions needed to be made and adhered to in every state. It was not sufficient to simply give the same test. It needed to be given at the same time. However, districts choose the starting day for schools, so giving the test to all schools on the same day means there will be a different number of learning days prior to the assessment, and giving the test after a specific number of school days had occurred could lead to security concerns with tests being given on different days across districts. Broadening the assessment window and recommending the number of learning days prior to assessing

helped alleviate that concern. Agreeing on an accommodations policy was also necessary, albeit one of the more difficult discussions among states.

PARCC struggled with comparability of administration platform as more states and districts in this consortium took a paper version of the test. In 2015, nearly 5 million students took the PARCC assessments, 81 percent on computer. Analyses by several states, including Illinois, Maryland, and Ohio, indicated that students did better on the paper version than on the computer version. A subsequent research study (Backes & Cowan, 2018) found mode effects of about 0.10 standard deviations in math and 0.25 standard deviations in ELA, which amounts to up to 5.4 months of learning in math and 11 months of learning in ELA in a single year. Interestingly, this mode effect was cut in half the second year of testing and was almost nonexistent by year 3. Possible reasons for these effects include unfamiliarity with devices, scrolling through reading passages versus flipping back and forth between pages in a booklet, and technology-enhanced items that cannot be fully replicated on paper. All of these factors affect the comparability of results, not just across states but within states, when students take the test on different modes.

### *State Versus National Assessments*

Since 2003, the National Center for Education Statistics (NCES) has released reports that map state proficiency levels onto the NAEP scale. Using an equipercentile linking approach, researchers match the percentage of students reported in the state assessment to be meeting the standard in each NAEP grade and subject to the point on the NAEP achievement scale corresponding to that percentage. They can thereby determine the NAEP equivalent of the state proficiency cut score. Next, a determination is made as to which NAEP performance level best matches the proficient level in each state. In more recent years, the match has been done at the school level. For example, if a state reports that 70 percent of the students in fourth grade in a given school are meeting their math achievement standard and 70 percent of the students in the NAEP achievement distribution in that same school are at or above 229 on the NAEP scale, then the best estimate from that school's results is that the state's standard is equivalent to 229 on the NAEP scale (see Figure 5-2). Results are then aggregated over all schools participating in NAEP in the state to provide an estimate of the NAEP scale equivalent of the state's threshold for its standard. Although not every school is assessed by NAEP, the sampling is done such that generalizations can be made to the entire state and standard errors are reported.<sup>2</sup>

The comparability results are reported and interpreted only at the aggregate state level. Additionally, only the proficient score is mapped, although, theoretically, additional cut points could be mapped. On the surface, such a broad comparison appears valid. However, a deeper dive into the warrants made by such a linking is needed.

A second type of state versus national assessment is conducted at the Stanford Education Data Archive. They use the state accountability data as well as NAEP data to conduct finer-grained analyses of issues like gender gap down to the district level. Their data set currently runs from 2008–2009 through 2014–2015. Some of their findings

---

<sup>2</sup> Additional details can be found at <http://nces.ed.gov/nationsreportcard/studies/statemapping>.

## Illustration of Mapping

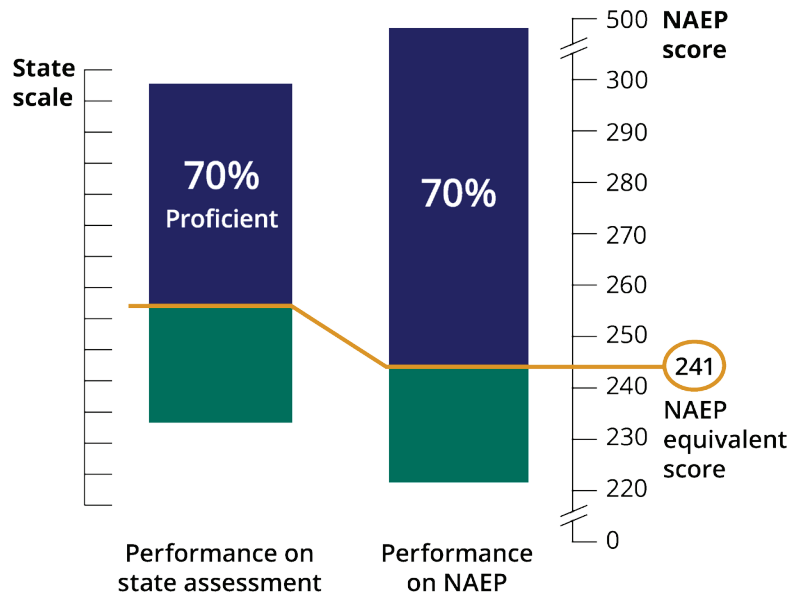


FIGURE 5-2 Illustration of mapping a state cut score onto the NAEP scale.

include that the average school district has no gender achievement gap in math but has a gap of roughly 0.23 standard deviation in ELA that favors girls. Both math and ELA gender achievement gaps vary among school districts but math gaps tend to favor males more in socioeconomically advantaged school districts and in districts with larger gender disparities in adult socioeconomic status. These two variables explain about one-fifth of the variation in the math gaps. However, they found little or no association between the ELA gender gap and either socioeconomic variable and can explain virtually none of the geographic variation in ELA gaps (Reardon, Fahle, Kalogrides, Podolsky, & Zárate, 2018).

**Purpose** The goals of state assessment systems are very different from those of NAEP. State assessments are specifically designed to be used in school accountability programs while NAEP is intended to be a snapshot of the performance of each state and the nation as a whole. Even though few state assessments are high stakes for students, the students do receive individual report cards that are sent home to their parents. Their teachers also understand that state assessments can affect them either directly, through a connection with performance reviews, or indirectly, through school ratings. This difference could affect the preparation teachers give to students prior to the testing window. Going into the assessments, students know that they will not receive scores on NAEP. Teachers also know that their school will not receive any feedback on student performance. Theoretically, then, students could approach the tests with differing levels of motivation to persevere on the more difficult items.



**Content** NAEP is not built to any particular set of content standards but rather to a framework determined by subject-matter experts and practitioners working for the National Assessment Governing Board (NAGB). NAEP frameworks provide the blueprint for the content and design of each NAEP assessment. In order to measure trends in student performance, NAEP frameworks are designed to remain stable for as long as possible; however, the frameworks are revisited approximately every 10–15 years to be responsive to changes in national standards and curricula. The current math and reading frameworks were published in 2009.<sup>3</sup>

In 2015, the NAEP Validity Studies Panel released a report of an alignment study conducted between the NAEP frameworks in mathematics and the CCSS at grades 4 and 8. The researchers found “reasonable agreement” overall but also some areas of fourth and eighth grade math where there was less of a match (Daro, Hughes, & Stancavage, 2015).

Specifically, the study found that 79 percent of NAEP items in fourth grade math assessed content included in the CCSS at grade 4 or below. However, the match rate was lower in some areas: 47 percent for data analysis, statistics, and probability; 62 percent for algebra; and 68 percent for geometry. In grade 8 the link was stronger, with 87 percent of NAEP items assessing math included in the CCSS at grade 8 or below. However, the authors noted that 42 percent of the CCSS for grades 6, 7, and 8 were not being tested by any items in the 2015 NAEP item pool. Therefore, there are definite content differences between the states using CCSS in 2015 and NAEP. For those states that were not teaching to the CCSS, the link is unknown but presumably no better.

**Administration conditions** Administration conditions can vary significantly across the two types of assessments. NAEP reading and math will be administered digitally for the first time in 2019. Many state assessments moved online years ago. As discussed in the previous section, there can be differences between online and paper versions of an assessment.

Students are given 60 minutes to take the NAEP items that have been selected for them. It is, therefore, a much shorter test than many state assessments, and each student only takes one subject. Because only a handful of students are selected to take NAEP in each sampled school, the assessment is given in a small-group setting, which is different from the classroom setting of most state assessments.

In the past, there have been concerns about students being opted out of NAEP because they either had a disability or were in an English learner program. In March 2010, NAGB adopted a new policy to maximize the participation of students with disabilities and English learners. Matching instructions under NCLB, NAGB recommended that exclusion rates should not exceed 5 percent of all sampled students. In 2017, approximately 90 percent of students with disabilities were included in the assessment. The only English learners that should be excluded are those who have been in U.S. schools for less than 1 year and for which a translated form of the assessment is not available.

Accommodation policies differ in some respects between some state assessments and NAEP. NAEP only allows a translated form for students who have been in U.S.

---

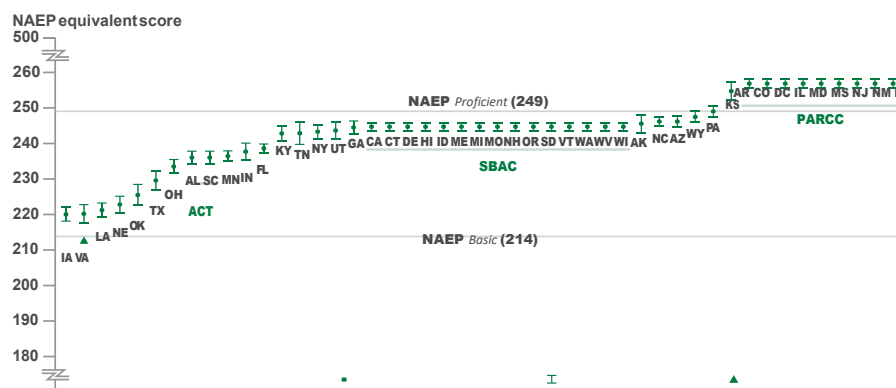
<sup>3</sup> Additional details can be found at <https://www.nagb.gov/naep-frameworks/frameworks-overview.html>.

schools for less than 1 year. After that, a word-to-word bilingual dictionary is provided. And these accommodations are not permitted on the reading or writing tests. Other accommodations match state accommodations such as extended time, directions read aloud, and test items read aloud for all but the reading test. The reading test also may not be presented to students in American Sign Language. Students are not permitted to have a calculator as an accommodation for the math or science tests. These accommodation policies are similar to those in some states but not all.

**Psychometric characteristics** Figure 5-3 shows the result of the state mapping done in 2015 for grade 4 mathematics. The claims made are that the proficient cut score set for PARCC in grade 4 math was slightly higher than the NAEP proficient cut score in the same grade and subject. Conversely, states such as Iowa, Louisiana, Nebraska, and Virginia set their proficient cut score at a level that was roughly equivalent to the NAEP basic level.

Although the psychometric characteristics of NAEP differ substantially from the state assessments, the claims made about NAEP seem reasonable. NAEP uses a matrix sampling approach to assessments, meaning different students receive different blocks of items, but all blocks are paired with one another to allow for an estimation of a full covariance matrix. That is, by giving a few students a few items, but by systematically spiraling those items in blocks and randomly sampling the students, inferences can be made about the full population. Because the performance on all items must be imputed based on multiple students taking a few items, background variables of those students are included in the estimation calculations. Multiple conditioned scores are produced and a sample of them is drawn to derive an ability estimate.

These sampled values are described as “plausible values” and only published in sufficiently sized aggregates, typically at the state and large-district levels. In this case, the only claims made are about the rigor of the benchmarks on the two assessments. Similarly, it is becoming common state practice for psychometricians to bring information about the percentage of students scoring at or above proficient on the NAEP grades 4 and 8 reading and math tests to the standard-setting workshops where cut



**FIGURE 5-3** NAEP scale equivalents of state grade 4 mathematics standards for proficient performance by state, 2015.

scores are placed on state assessments. Comparability is only discussed in terms of the expected level of rigor.

### **National Assessments Compared to International Benchmarks**

NCES ran a special study to link the NAEP scale to the TIMSS scale so that states could compare the performance of their students with that of students in other countries. First, it modified the NAEP assessment schedule so that eighth graders in all 50 states, the District of Columbia, and the U.S. Department of Defense schools could be assessed in mathematics and science in 2011. They were administered NAEP item booklets with some TIMSS items woven throughout. Then, nine states participated in the 2011 administration with a large enough sample to produce state-level results on TIMSS. They took TIMSS booklets that had NAEP items woven in. The NAEP results were used to link the two tests, and the TIMSS results from the nine participating states were used to validate the results. The design of the 2011 study allowed for the use of several different linking methods: statistical moderation, statistical projection, and calibration to predict TIMSS results for the U.S. states that participated in NAEP. All three methods produced similar results, so NCES chose to publish results from the statistical moderation analysis. "Statistical moderation aligns score distributions such that scores on one assessment are adjusted to match certain characteristics of the score distribution on the other assessment. In this study, moderation linking was accomplished by adjusting NAEP scores so that the adjusted score distribution for the public-school students who participated in 2011 NAEP had the same mean and variance as the score distribution for public school students in the TIMSS U.S. national sample. This allowed NAEP results to be reported on the TIMSS scale."<sup>4</sup> The analysis resulted in statements such as "Massachusetts and Vermont scored higher in science than 43 of the 47 participating education systems, while the District of Columbia scored higher than 14 education systems."

### ***Purpose***

NAEP, TIMSS, and PISA all purport to have different purposes and certainly serve different audiences. NAEP is a congressionally funded assessment that measures what U.S. students know and can do in various subjects across the nation, across states, and in some urban districts. It has multiple components dating back to 1969. TIMSS has measured trends in mathematics and science achievement at the fourth and eighth grades every 4 years since 1995. The goal is to get a snapshot of performance across multiple countries and to gauge progress over time. Finally, PISA is an international assessment that measures 15-year-old students' reading, mathematics, and science literacy every 3 years, emphasizing functional skills that students have acquired as they near the end of compulsory schooling. The age of 15 was chosen as it is the last age in which education is compulsory for most countries.

---

<sup>4</sup> Taken from [https://nces.ed.gov/nationsreportcard/studies/naep\\_timss/about\\_timss.aspx](https://nces.ed.gov/nationsreportcard/studies/naep_timss/about_timss.aspx).

### ***Content***

All three assessments are written to different test blueprints. NAEP and TIMSS are both curricula based while PISA is skills based, meaning it takes a broader approach to assessing student conceptual understanding. PISA assesses a different age of students than does NAEP or TIMSS, so the content is predictably different.

### ***Administration Conditions***

The NAEP-TIMSS linkage was designed with embedded items, meaning the administration conditions were exactly the same. However, that is not true for comparisons made between NAEP and PISA. As the next section describes in detail, NAEP tends to be more inclusive and offer more accommodations than PISA.

### ***Psychometric Characteristics***

From the student and teacher perspective, both national and international assessments have limited consequences. Therefore, the likelihood that teachers would engage in test-prep activities for either assessment is low, and students are likely to be equally (un)motivated for each. The linking study between NAEP and TIMSS used sound methodology, and the types of comparisons made appear reasonable. However, digging below the surface, TIMSS is based on specific science and math curricula that may be taught to students in the United States at different times. NAEP and TIMSS test specifications are not the same, so, even though the scores may be transferrable, assumptions about the level of knowledge and skills of a particular jurisdiction may not be.

## **Different International Assessments Compared**

There are three well-known international assessments used to rank-order countries based on student achievement. However, each of the assessments has differences in what and whom they assess. PISA focuses on reading, mathematical, and science literacy at age 15, rotating the emphasis each year; TIMSS assesses mathematics and science in grades 4 and 8; and the Progress in International Reading Literacy Study (PIRLS) focuses on reading at age 10. Few comparisons are made among the tests, with one exception. When PISA and TIMSS are given in the same year (e.g., 2003 and 2015), comparisons are made between the overlapping portions of each assessment. In 2003, comparisons were made regarding math achievement and in 2015 on science achievement.

Table 5-3 shows how each assessment differs on key features. Between PISA and TIMSS, it is important to note that they are given at two different ages (15 and 14, respectively) and are based on two different content frameworks. PISA is intended to be more general and is built around key concepts while TIMSS is curriculum driven, tests more specific knowledge and skills, and may show more differences in scores based on alignment with instruction.

**TABLE 5-3** Comparison of Key Features of Three International Assessments

	PISA	TIMSS	PIRLS
Full name	Programme for International Student Assessment	Trends in International Mathematics and Science Study	Progress in International Reading Literacy Study
Assesses	Reading, mathematics, science, problem solving	Mathematics and science	Reading
Age	15	10 and 14	10
Grade	9/10	4 and 8	4
Frequency	Every 3 years, since 2000	Every 4 years, since 1995	Every 5 years, since 2001
Last assessment	2018	2019	2016
When	Autumn	March–June	March–June
Purpose	Evaluates education systems by assessing to what extent students at the end of compulsory education can apply their knowledge to real-life situations and be equipped for society	Measures trends in math and science achievement	Measures trends in reading comprehension
Focus	Skills based	Curriculum based	Curriculum based
Parent organization	Organisation for Economic Co-operation and Development (OECD)	International Association for the Evaluation of Educational Achievement (IEA)	IEA
Countries	72 countries and economies in 2015	57 countries and 7 benchmarking entities in 2015	50 countries and 11 benchmarking entities in 2016
Test length	120 minutes, plus 35-minute background questionnaire	72 minutes at grade 4; 90 minutes at grade 8 plus 15-minute background questionnaire	80 minutes, plus 15-minute background questionnaire
Testing format in most recent year	Computer based	Computer based	Paper and pencil with an ePIRLS extension assessed online
Number of students assessed per country	More than 5,000	At least 4,000	About 3,500–4,000

### *Purpose*

Although the official purposes are listed differently, they appear to be used in similar manners. And student engagement, particularly in the United States, should not vary among the tests. Students are told that they are representing their country but will not receive their score.

### *Content*

There are clear differences in the content among the three assessments. First, they assess different, but overlapping, subject areas. Second, they assess different grades and ages and presumably align the content to be age and/or grade appropriate. Third, the tests developed by IEA are curriculum based while the tests developed by the Organisation for Economic Co-operation and Development are skills based, meaning that TIMSS and PIRLS test more specific knowledge and skills related to curriculum. Opportunity to learn should, therefore, have more of an impact on TIMSS and PIRLS scores than on PISA, which tests more general understandings.

### *Administration Conditions*

PISA is longer than TIMSS or PIRLS and has been assessed via computer. TIMSS was administered through paper booklets until 2019 when it moved to an electronic delivery of the assessment. PIRLS was also administered through paper booklets through 2016, although an optional technology literacy component was administered online. In 2021, PIRLS will transition to a fully online assessment. An additional distinction is that PISA is administered in the fall, while TIMSS and PIRLS are spring assessments.

TIMSS allows read-aloud accommodations only for the directions, as well as a calculator in grade 8 for all students. For students with disabilities who need a read-aloud accommodation, they may request words, phrases, or sentences be read aloud. For students requiring a calculator as an accommodation at grade 4, a school-supplied, four-function calculator is permitted. English learners may use a word-for-word dictionary for translation on TIMSS. Standard setting and presentation accommodations are provided. As of the 2015 assessment, PISA offers only limited accommodations for students with special needs, such as small-group settings, and their exclusion rate is higher. They do not permit extended time or allow large print, Braille, or even magnification. No assistance is provided for English learners. These differences could severely impact the comparability of TIMSS and PISA scores. PIRLS does not offer accommodations for English learners nor does it offer special forms for students with disabilities. Setting accommodations are allowed if typically used for other U.S. assessments.

All three of these policies are much less inclusive than typical U.S. state policies and result in more students being excluded from testing, a decision made at the school level. Indeed, exclusion rates across countries ranged from 0.0 percent in several smaller countries to 8.2 percent in the United Kingdom on the 2015 administration of PISA; the exclusion rate in the United States was 3.3 percent. For TIMSS, in that same year, student exclusion rates varied between 0.0 percent in eight countries to 6.8 percent in the United States. Again, assessing different populations could have a significant impact on comparisons of scores between the two assessments.



### *Psychometric Characteristics*

Typically, comparisons focus on how a country compares to others. One study from Germany found a strong correlation in mean scores by country across TIMSS and PISA (Kleime, 2016). In math, the coefficient of correlation is .923, indicating that 85 percent of the between-country variance in PISA mathematics literacy can be explained by TIMSS, and vice versa. Likewise, in science, the coefficient of correlation is 0.926, accounting for 86 percent of between-country variance. This indicates that the relative rankings could be compared, although the test takers and level of specificity of the content differ. It should be noted, however, that although individual-level correlations are unavailable due to the data-collection design, they would be expected to be much lower than these correlations between national averages.

Examination of claims made about country scores in analyses comparing TIMSS to PISA show that the statements are primarily about the rank ordering. For example, Wu (2009) writes

It is found that Western countries generally performed better in PISA than in TIMSS, and Eastern European and Asian countries generally performed better in TIMSS than in PISA. Furthermore, differences between the tests on two factors, content balance and years of schooling, can account for 93% of the variation between the differential performance of countries in PISA and TIMSS. Consequently, the rankings of countries in the two studies can be reconciled to a reasonable degree of accuracy.

These claims seem reasonable as they focus on overall trends and not specific comparisons of growth or absolute amount of knowledge demonstrated in the two assessments.

## CONCLUSION

There are many necessary conditions for full comparability of test scores. However, those conditions are not necessary to link scores; they are only necessary to validly interpret the results. That is, even though test scores can be linked, it does not mean that the interpretation is the same.

Determining the statement one wants to make about performance on the two tests determines the degree of comparability needed. For example, because the national-international comparisons focus on rank order of countries, the comparability rules are less strict. Conversely, colleges often assume that ACT and SAT scores are interchangeable simply because they can be concorded, even though the content differs, and they currently have different accommodation policies. Assuming interchangeable scores at the student level requires a higher degree of comparability than currently exists between the ACT and SAT.

Even when giving the same assessment on the same platform, states grappled with comparable administration times and conditions in the state consortia, where they wanted transportable scores. Full comparability is difficult to achieve, so it is important to understand the different characteristics and conditions of the assessments and to determine appropriate statements of comparison that can be made.

## REFERENCES

- Achieve. (2018). *Independent analysis of the alignment of the ACT to the Common Core State Standards*. Washington, DC: Author. Retrieved February 25, 2019, from <https://www.achieve.org/files/ACTReport.pdf>.
- Backes, B., & Cowan, J. (2018). *Is the pen mightier than the keyboard?: The effect of online testing on measured student achievement*. Washington, DC: American Institutes for Research. Retrieved February 27, 2019, from <https://caldercenter.org/sites/default/files/WP%20190.pdf?platform=hootsuite>.
- Bandeira de Mello, V., Blankenship, C., & McLaughlin, D. (2009). *Mapping state proficiency standards onto NAEP scales: 2005–2007: Research and development report*. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.
- Bandeira de Mello, V., Rahman, T., & Park, B. J. (2018). *Mapping state proficiency standards onto NAEP scales: Results from the 2015 NAEP Reading and Mathematics Assessments* (NCES 2018-159). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Retrieved February 10, 2019, from <https://nces.ed.gov/nationsreportcard/subject/publications/studies/pdf/2018159.pdf>.
- Braun, H. I., & Qian, J. (2007). An enhanced method for mapping state standards onto the NAEP scale. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales*. New York: Springer.
- College Board & ACT. (2018). *Guide to the 2018 ACT®/SAT® Concordance*. Retrieved February 12, 2019, from <https://collegereadiness.collegeboard.org/pdf/guide-2018-act-sat-concordance.pdf>.
- Daro, P., Hughes, G., & Stancavage, F. (2015). *Study of the alignment of the 2015 NAEP mathematics items at grades 4 and 8 to the Common Core State Standards (CCSS) for mathematics*. Retrieved February 20, 2019, from <https://www.air.org/sites/default/files/downloads/report/Study-of-Alignment-NAEP-Mathematics-Items-common-core-Nov-2015.pdf>.
- DePascale, C., Dadey, N., & Lyons, S. G. (2016). *Score comparability across computerized assessment delivery devices*. Washington, DC: Council of Chief State School Officers.
- Dorans, N. J. (2004). Equating, concordance, and expectation. *Applied Psychological Measurement*, 28(4), 227–246.
- Finn, B. (2015). Measuring motivation in low-stakes assessments. *ETS Research Report Series RR-15-19*. Retrieved February 13, 2019, from <https://doi.org/10.1002/ets2.12067>.
- Forte, E. (2017). *Evaluating alignment in large-scale standards-based assessment systems*. Washington, DC: Council of Chief State School Officers.
- Gewertz, C. (2019, April 19). Which states were using PARCC or Smarter Balanced in 2016-17? An interactive breakdown of states' 2016-17 testing plans. *Education Week*. Retrieved July 15, 2019, from <https://www.edweek.org/ew/section/multimedia/which-states-were-using-parcc-or-smarter.html>.
- Haniff, R. E. (2012). *The impact of timed versus untimed standardized tests on reading scores of third grade students in Title I schools*. Electronic Theses and Dissertations 2202. Retrieved from <http://stars.library.ucf.edu/etd/2202>.
- Holland, P. W. (2007). A framework and history for score linking. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales*. New York: Springer.
- Immekus, J., & Atitya, B. (2016). The predictive validity of interim assessment scores based on the full-information bifactor model for the prediction of end-of-grade test performance. *Educational Assessment*, 21(3), 176–195.
- Jia, Y., Phillips, G., Wise, L. L., Rahman, T., Xu, X., Wiley, C., & Diaz, T. E. (2014). *2011 NAEP-TIMSS linking study: Technical report on the linking methodologies and their evaluations* (NCES 2014-461). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.
- Kingston, N. M. (2009). Comparability of computer- and paper-administered multiple-choice tests for K-12 populations: A synthesis. *Applied Measurement in Education*, 22(1), 22–37.
- Klieme, E. (2016). *TIMSS 2015 and PISA 2015: How are they related on the country level?* (DIPF Working Paper). Retrieved February 24, 2019, from [https://www.dipf.de/de/forschung/publikationen/pdf-publikationen/Klieme\\_TIMSS2015andPISA2015.pdf](https://www.dipf.de/de/forschung/publikationen/pdf-publikationen/Klieme_TIMSS2015andPISA2015.pdf).
- Kolen, M. (2004). Linking assessments: Concept and history. *Applied Psychological Measurement*, 28(4), 219–226.

- Li, Y., Marion, S., Perie, M., & Gong, B. (2010). An approach for evaluating the technical quality of interim assessments. *Peabody Journal of Education*, 85(2), 163–185.
- Marion, S., & Perie, M. (2011). Some thoughts about comparability issues with “common” and uncommon assessments. Dover, NH: National Center for the Improvement of Educational Assessment. Retrieved February 2, 2019, from [www.nciea.org/sites/default/files/publications/Marion\\_Perie\\_Comparability%20paper\\_NCME\\_032511.pdf](http://www.nciea.org/sites/default/files/publications/Marion_Perie_Comparability%20paper_NCME_032511.pdf).
- Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects*. Princeton, NJ: Educational Testing Service.
- Nemeth, Y., Michaels, H., Wiley, C., & Chen, J. (2016). *Delaware system of student assessment and Maine comprehensive assessment system: SAT alignment to the Common Core State Standards*. Alexandria, VA: HumRRO. Retrieved February 25, 2019, from <https://www.doe.k12.de.us/cms/lib/DE01922744/Centricity/Domain/414/SATalignment.pdf>.
- NRC (National Research Council). (1999a). *Uncommon measures: Equivalence and linkage among educational tests* (M. J. Feuer, P. W. Holland, B. F. Green, M. W. Bertenthal, & C. Hemphill, Eds.). Washington, DC: National Academy Press.
- NRC. (1999b). *Embedding questions: The pursuit of a common measure in uncommon tests* (D. M. Koretz, M. W. Bertenthal, & B. F. Green, Eds.). Washington, DC: National Academy Press.
- Perie, M., Marion, S., & Gong, B. (2009). Moving towards a comprehensive assessment system: A framework for considering interim assessment. *Educational Measurement: Issues and Practice*, 28(3), 5–13.
- Pintrich, P. R. (1989). The dynamic interplay of student motivation and cognition in the college classroom. In C. Ames & M. Maehr (Eds.), *Advances in motivation and achievement: Vol. 6. Motivation enhancing environments* (pp. 117–160). Greenwich, CT: JAI Press.
- Pintrich, P. R., & DeGroot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, 82(1), 33–40.
- Reardon, S. F., Fahle, E. M., Kalogrides, D., Podolsky, A., & Zárate, R. C. (2018). *Gender achievement gaps in U.S. school districts*. Palo Alto, CA: Stanford Education Data Archive. Retrieved August 20, 2019, from <https://cepa.stanford.edu/content/gender-achievement-gaps-us-school-districts>.
- Reardon, S. F., Kalogrides, D., & Ho, A. (2018). *Linking U.S. school district test score distributions to a common scale* (CEPA Working Paper No. 16-09). Retrieved February 13, 2019, from <http://cepa.stanford.edu/wp16-09>.
- Roeber, E., Olson, J., Topol, B., Webb, N., Christophersen, S., Perie, M., Pace, J., Lazarus, S., & Thurlow, M. (2018). *Feasibility of the use of the ACT and SAT in lieu of Florida Statewide Assessments*. White paper commissioned by the Florida State Department of Education. Retrieved June 25, 2019, from <http://www.fldoe.org/core/fileparse.php/5663/urlt/ACTSATFSA.pdf>.
- Rogers, C. M., Lazarus, S. S., & Thurlow, M. L. (2014). *A summary of the research on the effects of test accommodations, 2011–2012* (Synthesis Report 94). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved June 25, 2019, from <https://nceo.umn.edu/docs/OnlinePubs/Synthesis94/Synthesis94.pdf>.
- Stephens, M., & Coleman, M. (2007). *Comparing PIRLS and PISA with NAEP in reading, mathematics, and science* (Working Paper). Washington, DC: National Center for Education Statistics, U.S. Department of Education. Retrieved February 22, 2019, from <http://nces.ed.gov/Surveys/PISA/pdf/compaper12082004.pdf>.
- Way, W. D., Davis, L. L., Keng, L., & Strain-Seymour, E. (2016). From standardization to personalization: The comparability of scores based on different testing conditions, modes, and devices. In F. Drasgow (Ed.), *Technology in testing: Improving educational and psychological measurement* (Vol. 2). Abingdon, UK: Routledge.
- Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (Council of Chief State School Officers and National Institute for Science Education Research Monograph No. 6). Madison: Wisconsin Center for Education Research, University of Wisconsin.
- Webb, N. L. (2007). Issues Related to Judging the Alignment of Curriculum Standards and Assessments. *Applied Measurement in Education*, 20(1), 7–25.
- Wu, M. (2019). A comparison of PISA and TIMSS 2003 achievement results in mathematics. *Quarterly Review of Comparative Education*, 39(1), 33–46.

# Comparability When Assessing English Learner Students

Molly Faulkner-Bond, *WestEd*

James Soland, *University of Virginia/Northwest Evaluation Association (NWEA)*

## CONTENTS

INTRODUCTION .....	150
Additional Assessments, Additional Decisions, and Additional Comparisons ..	150
The English Learner Population. ....	153
COMPARABILITY CONSIDERATIONS FOR ACADEMIC	
CONTENT ASSESSMENTS. ....	154
Accommodations .....	155
Academic Content Assessment for ELs with Disabilities. ....	156
Recommendations .....	156
COMPARABILITY CONSIDERATIONS FOR ASSESSMENTS OF ENGLISH	
LANGUAGE PROFICIENCY .....	159
The ELP Assessment Landscape .....	159
Interpreting Scores from Different ELP Assessments .....	160
ELP for ELs with Disabilities .....	163
Recommendations .....	163
ESTIMATING GROWTH IN ACHIEVEMENT AND LANGUAGE	
PROFICIENCY FOR ENGLISH LEARNERS .....	164
Reclassification and Opportunity to Learn. ....	165
Setting ELP and Reading Cut Scores .....	167
Construction of Vertical Scales .....	167
Estimating Growth .....	168
Recommendations .....	169
CONCLUSION .....	170
REFERENCES .....	171

## INTRODUCTION

Thus far, the chapters in this volume have addressed comparability issues that arise from efforts to compare the performance of individuals (Chapter 2, Comparability of Individual Students' Scores on the "Same Test") and groups (Chapter 3, Comparability of Aggregated Group Scores on the "Same Test") within (Chapter 4, Comparability Within a Single Assessment System) and across (Chapter 5, Comparability Across Different Assessment Systems) assessment systems. For any such comparisons, certain groups may face or introduce additional comparability considerations by virtue of their having special status or needs within those systems. One such group is that of English learner (EL) students—students designated through assessment as learning English in addition to academic content within schools. (Sireci and O'Riordan address a different subpopulation, students with disabilities, in Chapter 7, Comparability When Assessing Individuals with Disabilities, and Ercikan and Por address assessments in multilingual and multicultural contexts in Chapter 8, Comparability in Multilingual and Multicultural Assessment Contexts.)

The Every Student Succeeds Act (ESSA) of 2015 (the most recent reauthorization of the Elementary and Secondary Education Act [ESEA]) defines ELs as students whose difficulties in speaking, reading, writing, or understanding the English language may prevent them from accessing instruction or meeting challenging academic standards in a classroom where the language of instruction is English. In this chapter, we identify several decisions and test score uses that are specific to EL students in the United States and that introduce potential comparability issues for those who wish to make test-based generalizations or comparisons about this population of students, whether comparing them with one another or to non-EL students. Because EL students' education is governed by a different (and oftentimes more complex) set of test-based policies and systems than that of non-ELs, we begin this chapter by providing a brief overview of the EL context in the U.S. public school system.

### **Additional Assessments, Additional Decisions, and Additional Comparisons**

EL status is federally protected in the United States and is rooted in civil rights legislation. The underlying logic for the group's protected status is that students must not be excluded from equal educational opportunities on the basis of their race, color, or national origin—the last of which is proxied through home language (Pottinger, 1970; Smith, 1990; Williams, 1991). This protected status means that states and districts must have nondiscriminatory systems in place for identifying students who need this protected status. They also must offer services designed to help identified students overcome the vulnerabilities that make the protected status necessary. These services must include both supports to access mainstream instruction delivered in English, and language instruction to help EL students develop and meaningfully use their English to the point that the protections of EL status are no longer needed.<sup>1</sup> Implicit in this latter point is the idea that EL status is intended to be temporary: having expanded their English proficiency to the point that it no longer impedes their access to and meaningful

---

<sup>1</sup> Despite this mandate, it is important to note that the actual supports and services that students access may vary considerably in breadth and quality, across both contexts and individuals.

participation in instruction and opportunities, ELs are expected to exit the protected subgroup and leave behind the services and protections it is meant to provide. As we discuss at length in this chapter, identification for, and exiting from, this protected status is determined primarily by how ELs score on assessments of language and, in some settings, academic content.

Three specific test-based decisions that affect ELs are illustrated in Figure 6-1, which shows a roadmap for how ELs move through the typical U.S. school, including the assessment on which each step is based. There are three main decision points for a potential EL. First, when any students initially enter the school system, they are administered a home language survey (HLS), which is often a few basic survey questions about a student's language background, and the language used in the home. Second, students who are deemed to potentially need EL services on the basis of their HLS responses then take an English language proficiency (ELP) test (often referred to as a "screener") to better determine their English proficiency and their needs in terms of language support. Typically, students who score below a particular language proficiency cut score are then designated as ELs. Once a student is identified as an EL, states are required provide them with the services described above (i.e., services to support access to mainstream instruction delivered in English, and language instruction to help EL students develop and use their English proficiency for classroom learning). States also must annually assess EL students' ELP. These scores are used both to track

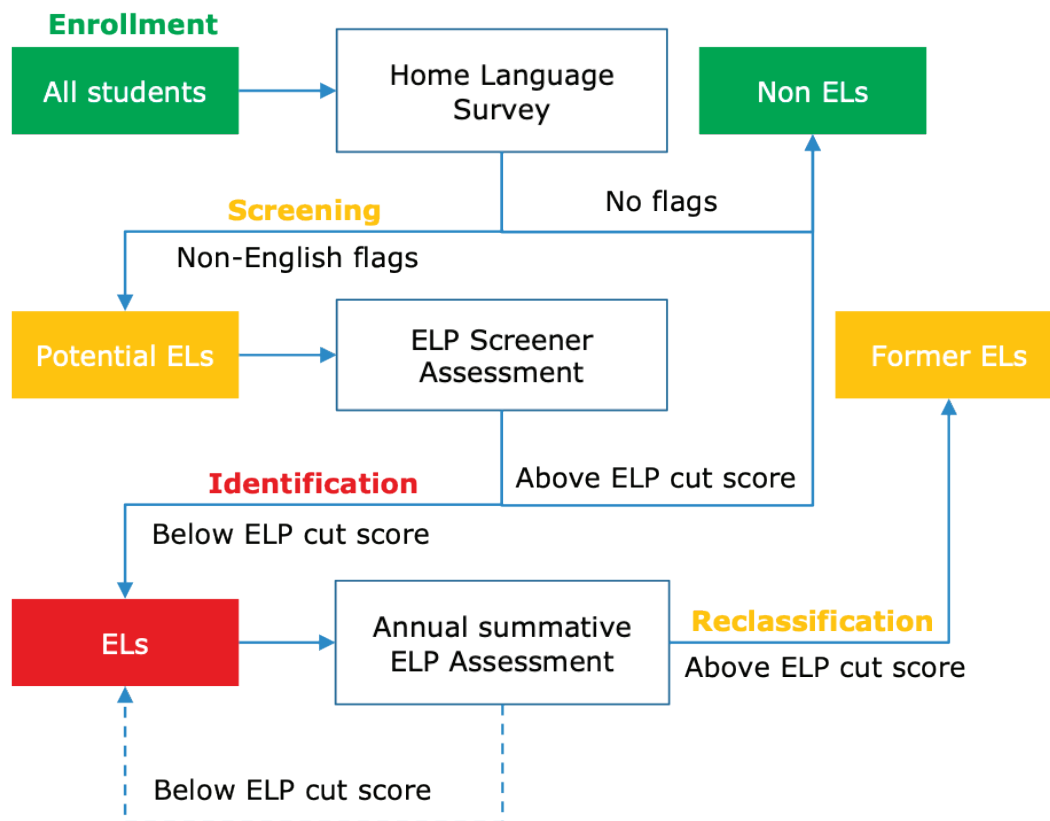


FIGURE 6-1 Identification and reclassification life cycle for an English learner.



individual students' progress on ELP and for systems-level accountability. Once ELs are judged to have attained proficiency based on their scores on ELP (and oftentimes other) tests, they take the third major step in the EL life cycle: reclassification as "fully English proficient." Upon being reclassified, ELs typically no longer have access to whatever additional supports they were receiving related to their academic and language needs.

For all of the steps described above, states—and in some cases, districts within states—have considerable autonomy in developing or selecting measures and criteria. Home language surveys are generally ad hoc by district rather than standardized. For screener tests, states may design or select one or more assessments to use and also may set their own cut scores to decide which students will and will not be identified as ELs. As we discuss in more detail below, assessments to measure ELP and related cut scores to determine proficiency also vary across states. And, finally, reclassification criteria differ across states (and even districts in some cases), both in terms of cut scores and in terms of which tests and scores are even considered. Thus, each decision in the EL life cycle is based on different instruments, constructs, cut scores, and decision rules across settings, introducing comparability issues as fundamental as who gets labeled as an EL in the first place.

Research also shows that all three assessment-based decisions have major implications for ELs' opportunity to learn. As noted above, EL status is typically accompanied by academic supports like English language development classes, modified instructional content, instruction from specially prepared teachers, and the aforementioned annual assessments of ELP (Pompa & Villegas, 2017; Robinson, 2011; Umansky & Reardon, 2014). At the time of school entry, students who need language supports to be successful in the classroom could be at a disadvantage if they are not identified as EL by the HLS, screener, or both. Conversely, there is evidence that EL status can be stigmatizing (Dabach, 2014) and difficult to exit (Robinson, 2011), which means a student whose need for language supports is borderline (or who is identified as EL in error) might actually be disadvantaged through identification. There is also considerable evidence that reclassification has implications for opportunity to learn. Once a student is reclassified, language supports diminish or disappear entirely. Research using rigorous causal methods suggests the timing of reclassification is a major key to ELs' long-term academic success (Robinson, 2011; Robinson-Cimpian & Thompson, 2015). Therefore, measuring ELP reliably for ELs and setting reasonable cut scores on ELP assessments is consequential for this group of students.

Perhaps based on the implications of EL status for opportunity to learn, federal law holds schools accountable for how quickly ELs move toward English proficiency in hopes of reducing the time that elapses between identifying a student as an EL and reclassification. The reauthorization of ESEA as the No Child Left Behind Act (NCLB) in 2001 created a major shift in federal accountability for ELs: for the first time, states were required to annually measure and report ELs' overall English proficiency, as well as their proficiency in the subdomains of listening, speaking, reading, and writing (Abedi, 2004). Furthermore, schools were being held accountable not only for ELs' achievement, but also for their progress toward being proficient in English as measured by those standardized ELP assessments (Abedi, 2004; Cook, Linn, & Jung, 2012). Under ESSA, the core assessment and accountability requirements for ELs remain (CCSSO, 2016), but a requirement has been added for states to standardize how ELs are reclassified. Ultimately, while the types of assessments required under ESSA have

not shifted, the ways they are used to hold schools accountable have expanded. These accountability requirements therefore mean test score comparability issues have implications not only for students but also for the schools that serve them.

### **The English Learner Population**

The population of students identified through the test-based system described above is large and diverse. By the numbers, more than 5 million students are designated as English learners in public school, accounting for more than 10 percent of the K–12 population, which is an increase of more than 50 percent in the past decade (Cook, Boals, & Lundberg, 2011). While federal law tends to treat EL status as binary, this subpopulation is diverse in terms of language and academic background, which can complicate measurement and comparability of relevant test-based criteria.

On one hand, ELs generally share certain commonalities. For example, a considerable majority of ELs are young Spanish speakers born in the United States and raised in underresourced communities (NASEM, 2017) where teachers may be less than well prepared to teach ELs (Gándara & Santibañez, 2016). On the other hand, incredible diversity exists within the population. ELs are found in every grade and state, and there are hundreds of languages represented and used by ELs and their families across the United States (Aguirre-Muñoz & Boscardin, 2008; Dabach, 2014; Pompa & Villegas, 2017). Within the population of Spanish speakers, considerable dialectal diversity exists (Solano-Flores & Li, 2009). The exact composition of the EL population also varies from setting to setting; for example, some districts may have large numbers of newly arrived, displaced refugees, whereas others have well-established communities that have been part of the district for generations. Academically, ELs enter the school system at different points in their schooling and language development, and at different times of the school year (Allard, 2016).

Like non-ELs, ELs also may have learning or cognitive disabilities. Recent data suggest that ELs mirror the general population in terms of which disabilities are most common. Specific learning disabilities are the most prevalent disability classification in both groups (reported for roughly half of all students with disabilities and roughly 40 percent of ELs with disabilities), followed by (in rank order) speech or other language impairments, other health impairments, autism, and intellectual disabilities (Wu, Liu, Thurlow, & Albus, 2019). Unlike non-ELs, the distinction for EL students between language development and language-based disability can be difficult to ascertain clearly and may create challenges for identifying students for EL services, special education services, or both. These differences complicate how test makers define and measure constructs like English language development and proficiency, as well as how ELs with disabilities should best be included in large-scale assessments.

Both the policy-driven nature of EL status and the diversity of the EL subpopulation introduce many of the comparability issues we go on to discuss in the remainder of this chapter. As we have shown above, the question of whether a student should be an EL at all may be answered differently in different settings. On top of these test-based differences in the population itself, we spend the remainder of this chapter discussing other comparability issues that arise for students who are identified as ELs. Scores on achievement tests cannot always be straightforwardly compared between EL and

non-EL students, given potential confounds between language and academic ability (Abedi, 2006). Furthermore, many ELs take achievement tests using accommodations that complicate comparisons if not properly addressed, and ELs can also be part of other subgroups like students with disabilities that necessitate additional accommodations. Within the EL subgroup, interpreting ELP scores can be influenced by differences in how tests are developed and scored, how states weight various subscores, and even how the construct of language proficiency is operationalized across measures. Finally, using scales to estimate and compare growth for ELs (including comparing achievement growth estimates to those for non-ELs) is complicated by the shifting nature of the EL subgroup. In each section that follows, we discuss one of these challenges to score comparability. We also present associated recommendations for policy, practice, and research.

### COMPARABILITY CONSIDERATIONS FOR ACADEMIC CONTENT ASSESSMENTS

Prior to NCLB, states could (and routinely did) exclude ELs from academic content assessments often with the rationale that, because ELs had not yet mastered English, their test scores were unlikely to be valid or meaningful (Abedi & Lord, 2001; Martiniello, 2008). While NCLB (and ESSA after it) did not necessarily dispute this notion, the law operated on a different logic, which persists to this day: that ELs who are excluded from instruction and assessment are not typically placed in similarly rigorous or high-quality alternative environments. By forcing states and districts to report and reckon with the subgroup's academic performance, the law sought to motivate agencies to take these students' education more seriously and include them in grade-level instruction (Faulkner-Bond & Forte, 2016; Thurlow & Kopriva, 2015).

Although this rationale of inclusion may be valid and important from a policy perspective, it does create measurement challenges over how to ensure the validity of these students' scores given their ongoing language development (Sireci & Faulkner-Bond, 2015).<sup>2</sup> Given that ELs are, by definition, still developing their proficiency in English, it is likely that language may affect their ability to engage with the assessment and demonstrate their knowledge and skills on the construct being measured (AERA, APA, & NCME, 2014; Sireci & Faulkner-Bond, 2015; Solano-Flores, 2014). For non-ELs, we assume that language plays no meaningful role in their performance. For an EL, however, we often cannot be sure whether an incorrect answer is due to language proficiency, content knowledge, or both (Abedi & Lord, 2001; Martiniello, 2008). For older ELs who have been in the system for several years, opportunity to learn also may be a factor affecting performance and interpretation given research suggesting that EL status is often associated with diminished access to rigorous courses (Umansky, 2016).

Furthermore, evidence suggests that the exact role that language does play, and the ways that that role might be managed, likely also vary as a function of a student's language proficiency. Wolf and Leon (2009) found, for example, that items on a science

---

<sup>2</sup> Although experts have advocated for more nuanced ways to report and account for the performance of ELs on academic content assessments (e.g., by weighting scores based on students' level of English proficiency) (Cook et al., 2012; Working Group on ELL Policy, 2010), the U.S. Department of Education has so far declined to grant states permission to pursue this type of approach.

assessment exhibited differential item functioning within the EL subgroup, showing bias for ELs with lower language proficiency compared to those with higher ELP. This finding suggests that the role of language in content assessments varies as a function of language proficiency, setting up comparability concerns even within the EL population. As such, comparability issues may arise within states whether one is comparing ELs to non-ELs or ELs to one another.

### Accommodations

Traditionally, the response to this comparability concern has been to provide ELs with accommodations to ensure they can access assessment content and demonstrate their knowledge and skills. As discussed in Chapter 7, *Comparability When Assessing Individuals with Disabilities*, ideally implemented assessment accommodations should (1) provide a “differential boost” for students who need them (Abedi, 2009; Lane & Leventhal, 2015; Sireci, Scarpatti, & Li, 2005), (2) not alter the focal construct of the assessment or item, and (3) not affect the difficulty of the content being measured (Thurlow & Kopriva, 2015). For EL students, this generally translates to a goal of reducing the potential for construct-irrelevant language to hinder students’ performance, while keeping the construct-relevant aspects of the assessment (linguistic and substantive) unchanged. In cases where these standards are met, scores from accommodated and nonaccommodated examinees may be considered comparable, and aggregating them may be justified (Abedi, 2016). When evidence for any of the conditions above is lacking or inconclusive, best practice dictates that the scores not be aggregated or treated as comparable.

To date, there are relatively few accommodations for ELs that meet the standards above for comparability and score aggregation. Across several syntheses of research on the effectiveness of accommodations for ELs on content assessment (Abedi & Ewers, 2013; Kieffer, Rivera, & Francis, 2012; Li & Suen, 2012; Pennock-Roman & Rivera, 2011), the consensus is that customized language glossaries either in standard English or with translations to a home/community language, particularly when offered in combination with extra testing time, are the only accommodations that are effective for EL students without compromising the validity of their responses. The same reviews also recommend simplified language as an effective option, though they differ in whether they recommend this as an accommodation (Abedi & Ewers, 2013) or as a consideration for test construction (Kieffer et al., 2012).

Perhaps in response to the findings above, the approach to accommodating ELs and all students on academic content assessments has shifted in recent years (Shyyan et al., 2017), in ways that ultimately are beneficial for comparability. Specifically, based on recommendations and guidance from experts in the field (Abedi & Ewers, 2013; Solano-Flores, 2012), the Race to the Top assessment consortia (Smarter Balanced and the Partnership for Assessment of Readiness for College and Careers [PARCC]) adopted several supports on their assessments that are available to all students without needing special permission. These include English glossaries, English dictionaries, a thesaurus, instructions read aloud in English, and certain reading supports that change the formatting of presented text (e.g., by increasing the text size, or presenting it line by line rather than in a block) (PARCC, 2017; Smarter Balanced Assessment Consortium,

2014). Extended or unlimited time options are also offered as first-level accessibility tools, meaning they must be specifically requested for a given student, but do not rise to the level of being considered an accommodation (PARCC, 2017; Smarter Balanced Assessment Consortium, 2014). Other EL-specific supports at this tier include targeted word-level translations, translated or dual-language test forms, and options for alternative means of expressing one's answers (e.g., the use of a scribe or speech-to-text technology). By moving accommodations with a solid research basis into the category of universal access features, the consortia fundamentally changed how many ELs experienced assessment and also reduced the number of accommodations that may introduce comparability challenges for EL-to-English-only comparisons.<sup>3</sup>

### **Academic Content Assessment for ELs with Disabilities**

EL students with disabilities may need additional accommodations to ensure they can interact with tests of academic content. For this kind of assessment, accommodating ELs with disabilities functions largely as an extension of the "regular" accessibility process for all ELs, with additional supports made available to EL students with disabilities to complement their language-related tools. Ideally, such students will be assigned accommodations that work together to address both needs simultaneously, rather than via separate teams or processes, to ensure that students are not inundated with duplicative or even conflicting accommodations during the assessment (Liu, Ward, Thurlow, & Christensen, 2015).

Given the intersection of their status as ELs and students with disabilities, ELs with disabilities may also have two reference groups of interest for comparability, namely, students with disabilities or ELs. For such comparisons, the same comparability logic from above would also presumably apply: if any ELs with disabilities can access academic content assessments using only accessibility features or accommodations that do not alter item difficulty or content, then such students' scores could feasibly be considered comparable to those from students without these accommodations and aggregated for interpretation. When these conditions are not met, scores from this population will not be comparable to scores from ELs without disabilities, English-only students with disabilities, or both.

### **Recommendations**

Even with appropriate accommodations, it is possible and even likely that scores between some ELs and non-ELs on the same content assessment may still lack comparability. This will likely be more salient for individual-level comparisons, and most applicable for ELs with the lowest levels of ELP compared to higher-ELP EL students or English-proficient students. In these cases, students' low levels of ELP may still affect their performance, even when effective accommodations are provided. Thus, individual-level comparisons should be made with caution.

---

<sup>3</sup> Although many states have shifted back toward state-specific assessments in recent years, this tiered approach to accommodations and accessibility has persisted and even increased as more and more tests are offered digitally, where technology increases the ease and efficiency of offering both universal and tailored supports on an individual basis.



In addition, while it is important to ensure that all students, including ELs, have access to the accommodations and supports they may need to demonstrate their knowledge and skills, it is equally important to avoid a scenario in which unnecessary access issues are created through a lack of attention either to the importance (or nonimportance) of language in the construct being measured, or to the particular supports that EL students may need, when tests are being built. Accommodations should ideally be used to scaffold students' access to content rather than to correct for errors or inaccessibility that were unwittingly introduced in the test construction process. To this end, there are several steps that test users and developers may take to minimize the potential for comparability issues when making within-state comparisons involving ELs on academic content assessments.

### *Articulate or Consider the Role of Language in the Content Domain*

The current generation of content standards used by states (e.g., the Common Core State Standards [CCSS], the Next Generation Science Standards, and the state-specific variants of these in use around the country) often explicitly acknowledges disciplinary language uses in their articulation of the content domains (Frantz, Bailey, Starr, & Perea, 2014). This effort to explicitly articulate when, where, and why language does and does not matter relative to the content that surrounds it is an improvement over previous generations of content standards, which were largely silent on the role of disciplinary language (Bailey, Butler, & Sato, 2007). These advances also help send a clear signal to educators that language practices should be a part of their instruction and instructional planning if they wish to cover the full content of the academic standards. To the extent that test developers design blueprints that are reflective of the standards or domain a test is intended to measure, this inclusion of language uses within academic content standards should help to ensure that the language uses on the test form itself are appropriately reflective of the construct-relevant language uses of the domain (Avenia-Tapper & Llosa, 2015).

### *Be Aware of Language During Item Development*

A growing field of research has identified specific item features and language constructions that appear to differentially affect ELs on assessments of academic content (Kachchaf et al., 2016; Martiniello, 2009; Noble, Rosebery, Suarez, Warren, & O'Connor, 2014; Wolf & Leon, 2009). Specific features that have been identified across studies include the forced comparison item format ("which of the following...") (Kachchaf et al., 2016); low-frequency, nontechnical vocabulary (i.e., vocabulary that is not directly content related and that is used infrequently in other texts or settings that students might encounter at home or in school) (Kachchaf et al., 2016; Martiniello, 2008; Wolf & Leon, 2009); and long or complex item stems (Kachchaf et al., 2016; Martiniello, 2009). Several studies also have found that the use of visuals, diagrams, or schematics can make items more accessible for ELs (Kachchaf et al., 2016; Martiniello, 2009; Solano-Flores, Wang, Kachchaf, Soltero-Gonzalez, & Nguyen-Le, 2014). All of these findings might be incorporated into item-writing guidelines (Noble, Rosebery, Kachchaf, & Suarez, 2016; Solano-Flores, 2014), item review protocols, and test blueprints (e.g., "no more than X% of the test's items should use forced comparison constructions").



### *Include and Oversample ELs in the Test Validation Process*

Solano-Flores (2014) in particular has advocated for oversampling of ELs in pilot and field test samples, as well as conducting cognitive laboratories with EL students to understand how they are interpreting items. The purpose of any such steps would be to try to ensure sufficient samples sizes to conduct the necessary analyses and comparisons to identify instances in which language acts as a source of construct-irrelevant variance in EL students' performance. By identifying such instances during the test construction process, test developers can preemptively revise and correct items and blueprints as needed to reduce the need for accommodations once a test becomes operational (Faulkner-Bond & Forte, 2016; Sireci & Faulkner-Bond, 2015; Solano-Flores, 2014).

### *Broaden the Options for Engaging with Content and Conveying Knowledge and Skills*

Kopriva and colleagues have argued for vastly expanding the nature and variety of methods offered to students during both instruction and assessment to demonstrate their knowledge of content skills (Kopriva, 2014; Thurlow & Kopriva, 2015). Their ONPAR (Obtaining Necessary Parity Through Academic Rigor) assessment system offers ELs and non-ELs alike multiple methods of receiving and expressing content that largely remove language from the equation and instead allow students to engage directly with mathematic and scientific concepts through dynamic, multisemiotic features (Kopriva & Wright, 2017; Logan-Terry & Wright, 2010). For example, students may demonstrate the effects of friction on an object's speed by selecting from a series of images and animations that represent—often without any language at all—the object slowing down, speeding up, being unaffected, etc. Although such assessments might not provide valid information about students' ability to use disciplinary language, the authors argue that this approach ultimately would allow ELs (and all students) opportunities to demonstrate their content understanding without having to worry about language clouding their performance or our interpretations thereof.

On this note, it is worth acknowledging an important assumption embedded in considerations of comparability issues for ELs on content assessments: that language plays no role in the performance of non-EL students. While intuitive, this is indeed an assumption, and not something that is tested empirically for any large-scale academic content assessments of which we are aware.<sup>4</sup> There is some evidence to suggest that this assumption may be problematic: Carroll and Bailey (2015) found, for example, that when they administered a state's ELP assessment to non-EL students for research purposes, anywhere from one-fifth to one-third of the sample (depending on the decision rule used) did not earn scores high enough to achieve reclassification. This suggests that the non-EL population includes students whose proficiency falls below the stated threshold for ELP on the ELP assessment. In light of these findings, we emphasize that comparisons of EL and non-EL performance on content assessments are asymmetric, in the sense that we know more about ELs' language abilities than we do about that of

---

<sup>4</sup> An extensive synthesis of studies addressing validity and psychometric issues in the assessment of ELs (Lane & Leventhal, 2015) identified only two studies that looked at factorial invariance for ELs and non-ELs on the same assessment.

students who are never identified as English learners. Thus, the true scale and extent of any comparability issues arising between these two groups may not be fully known until or unless the language knowledge and skills of all students are measured (Carroll, 2012) or explored through methods that attend to the structure of responses from different subgroups (Lane & Leventhal, 2015; Sireci & Wells, 2010).

### COMPARABILITY CONSIDERATIONS FOR ASSESSMENTS OF ENGLISH LANGUAGE PROFICIENCY

In addition to the academic content assessments administered to the general population, all EL students (and only EL students) in all grades (K–12) must participate annually in assessments to measure their ELP. As noted in the Introduction, these scores are used for consequential decisions at both the student level, where they are used to support reclassification decisions (Carlson & Knowles, 2016; Carroll & Bailey, 2015; Kieffer & Parker, 2016; Robinson, 2011), and the school and district levels, where they are used for accountability purposes to evaluate program and instructional quality (Robinson-Cimpian & Thompson, 2015; Umansky & Reardon, 2014). In this section we review several comparability challenges that arise from these assessments and their uses.<sup>5</sup> For this discussion, all comparisons are among EL students (since non-ELs do not participate in these assessments) and generally concern cross-state or cross-district comparisons.

#### The ELP Assessment Landscape

In contrast to the content assessment landscape, where interstate collaboration was not a norm prior to the era of the CCSS and *Race to the Top*, states have chosen and been encouraged to work together on developing ELP standards and assessments since NCLB implementation began in 2002. Between assessments developed by government-funded state consortia and off-the-shelf assessments from testing companies, the number of ELP assessments in use nationwide has never exceeded about a dozen, at least in terms of item pools and test blueprints (Abedi, 2007) (though states have had options to tailor forms, cut scores, or reporting to their own context).

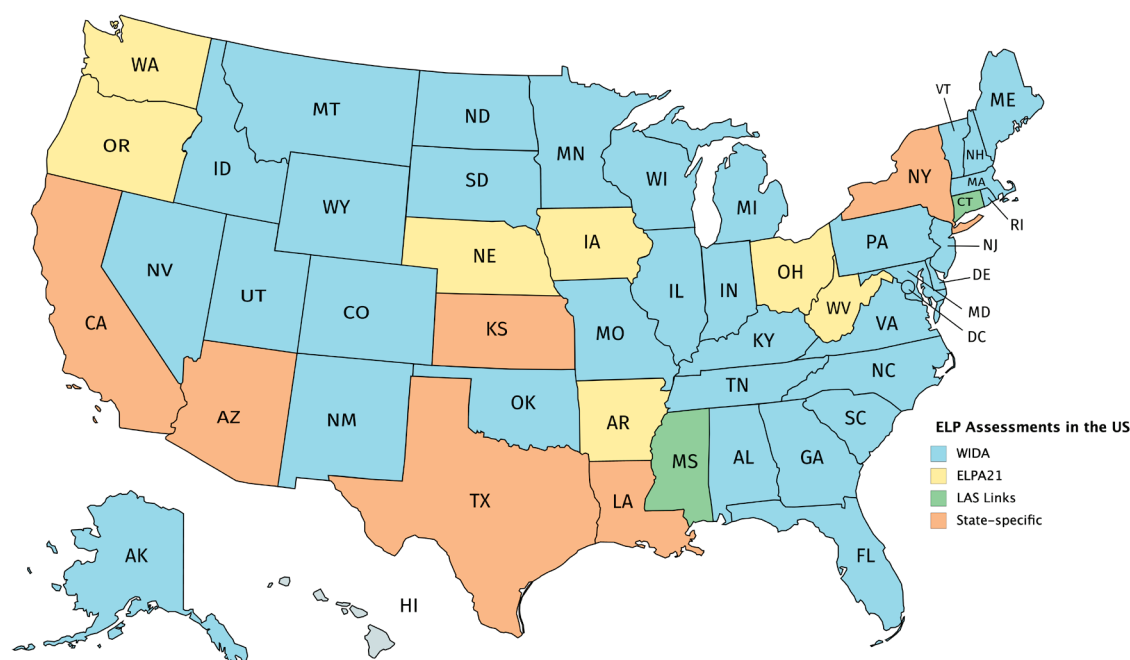
Despite ESSA's expansion of state-level discretion, the past decade has been marked by further consolidation of the ELP assessments used among states. Today, only nine different ELP assessments are in use across the 50 states and the District of Columbia. As shown in Figure 6-2, the majority of states (35, plus the District of Columbia) use the WIDA<sup>6</sup> assessment,<sup>7</sup> while seven states use the English Language Proficiency

---

<sup>5</sup> As a note, we focus solely on summative annual ELP assessments here. Due to space constraints, we do not address screener assessments or home language surveys, although—as noted in the “Introduction” in this chapter—these instruments also can be consequential for EL students.

<sup>6</sup> The WIDA consortium was created through an Enhanced Assessment Grant to the states of Wisconsin, Delaware, and Arkansas in 2003. Initially, “WIDA” was an acronym incorporating the names of the three states. The consortium subsequently changed its name to “World-Class Instructional Design and Assessment” as its membership grew to include more states. Eventually, in 2016, the consortium dropped the underlying title and now simply goes by “WIDA” (not unlike the SAT).

<sup>7</sup> Four additional states use WIDA's alternate assessment for ELs, despite using a different assessment for their primary ELP assessment.



**FIGURE 6-2** ELP assessments in use across the United States.

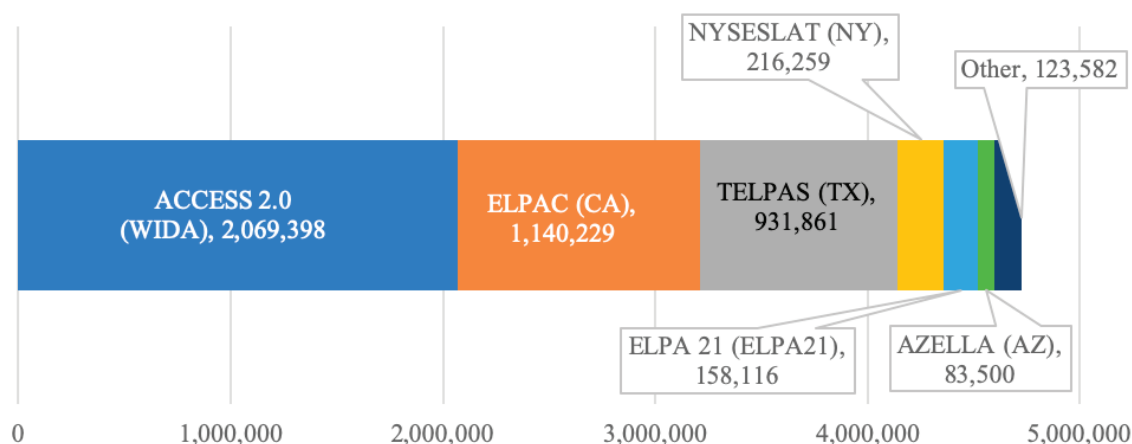
Assessment for the 21st Century (ELPA21) assessment and two states use the LAS Links.<sup>8</sup> The remaining six states—Arizona, California, Kansas, Louisiana, New York, and Texas—use state-specific assessments. By the numbers, WIDA’s ACCESS for ELLs assessment (ACCESS) is the most widely used (as shown in Figure 6-3, it was administered to more than 2 million EL students in the 2018–2019 school year), followed by the state-specific assessments for California and Texas, respectively.

### Interpreting Scores from Different ELP Assessments

On one hand, the use of common ELP assessments is a boon to comparability. For the many students taking the WIDA assessment, for example, test scores are based on equivalent constructs, scoring models, and proficiency descriptors. On the other hand, even with only nine assessments to compare, challenges remain for those who wish to compare the progress of ELs across states or nationally. Some of these challenges are similar to those encountered for academic content assessments (see Chapter 2, Comparability of Individual Students’ Scores on the “Same Test,” and Chapter 3, Comparability of Aggregated Group Scores on the “Same Test”), while others arise from unique aspects related to how most ELP assessments are built and scored.

First and most straightforwardly, states that use the same ELP assessments still have the flexibility to set their own cut scores for proficiency. Although every state using the

<sup>8</sup> Initially, the “LAS” in “LAS Links” stood for “Language Assessment Scales”; however, the test’s owner (previously CTB/McGraw-Hill, now Data Recognition Corporation) no longer uses the underlying definition, and simply refers to the test as the LAS Links.



**FIGURE 6-3** Number of ELs participating in different assessments.

NOTE: The numbers in this figure reflect the most recent data available according to state and consortia websites.

WIDA assessment uses the same labels and the same thresholds for the assessment's six performance levels, states are free to decide which point on this shared scale is sufficient to consider a student ready for reclassification. Thus, for the most consequential use of an ELP score, states sharing common assessments actually cannot claim that reclassified students have scores in the same range on the test's scale.

For states that use different ELP assessments, additional comparability issues arise. As with academic content assessments, different ELP assessments are built based on different English language development standards, which vary in how they define and articulate the construct of interest (i.e., the English language knowledge skills necessary to support academic learning and performance). WIDA's standards, for example, are explicitly structured around the language of different academic content areas like science and social studies (WIDA Consortium, 2012), whereas the standards used by California and the ELPA21 states are more discipline agnostic in their language and instead provide ancillary materials demonstrating how the language standards relate to different content areas (California Department of Education, 2012; CCSSO, 2013). Thus, although all ELP tests produce, for example, a reading score, the construct and content underlying those scores may differ from setting to setting in important ways.

Second, in response to the language of both NCLB and ESSA, most ELP assessments produce separate scores for the four domains of reading, writing, listening, and speaking. These domains are measured through discrete subtests that may be administered in different ways or on different occasions. The scores are then combined in different ways to produce a composite score to represent overall language proficiency. In contrast to academic content assessments, which are generally shown to be strongly unidimensional, the construct of ELP is generally agreed to be multidimensional (Faulkner-Bond, Wolf, Wells, & Sireci, 2018). As a result, the way these scores are combined has profound implications for interpretations of student performance, including comparative interpretations.

Figure 6-4 shows how the nine assessments currently in use combine subscores to create their overall composites. As the table makes clear, the construct of overall English

	Reading	Writing	Listening	Speaking
	Percentage of Overall Proficiency Composite			
ACCESS 2.0	35	35	15	15
AZELLA*	32	30	18	20
LAS Links NYSESLAT TELPAS	25	25	25	25
ELPAC (1-12)	50 (Written Composite)		50 (Oral Composite)	
ELPA21 ELPT (LA) KELPA2	N/A – These assessments use profiles rather than a single composite score			

FIGURE 6-4 Overall score domain weights for different ELP assessments.

<sup>a</sup> Domain weights vary slightly across test forms for the AZELLA; this table represents averages across all grade spans.

proficiency is operationalized differently by different tests, creating comparability issues across states that use different instruments. Three of the seven tests (LAS Links, NYSESLAT, and TELPAS) combine the scores by either a sum or an unweighted average, while one test (ELPAC) combines two composite scores with equal weights. This approach suggests a compensatory interpretation of proficiency (Carroll & Bailey, 2015), wherein lower performance in some domains may be compensated to some degree by higher performance in other domains. This also means that students with different subscore profiles may earn the same overall score.

Two tests (ACCESS and AZELLA) use unequal weights across the domains. On these tests, the use of weights and composites suggest an effort to maintain but minimize compensatory effects by tipping the balance of the overall score toward domains believed (or shown) to be more important in predicting or supporting academic content achievement. The remaining three assessments (ELPA21, KELPA, and ELPT) do not create a composite score but rather make overall decisions based on a multivariate score profile, which is essentially a vector of the performance levels from the four domain subtests (e.g., 5-3-4-5) (Cai & Hansen, 2018). On these assessments, all of a student's domain scores must be in one of the two highest performance levels on each subtest to achieve proficiency.

Notably, too, some states use a combination approach, wherein a student's overall composite score must reach a certain threshold *and* other subtest scores also must be at or above a certain level. This type of conjunctive approach reflects a belief that students must have some minimum level of achievement in some or all domains to be considered proficient, and compensation across domains is not relevant or appropriate for making decisions (Carroll & Bailey, 2015). As these differences suggest, the meaning of a student's overall proficiency level may differ across ELP assessments as a result of these different methods.

### ELP for ELs with Disabilities

Several of the comparability issues surrounding ELP scores are even more complex for EL students with disabilities. First, one should note that the role and nature of accommodations is fundamentally different on ELP assessments than it is for content assessments. On content assessments, as discussed above, the general goal for ELs is to reduce the role of construct-irrelevant language, so that students may demonstrate their content knowledge and skills with minimal linguistic interference. On ELP assessments, by contrast, the language *is* the construct. Thus, accommodations are used only to remove barriers students may face in demonstrating their language knowledge and abilities to communicate meaningfully. In practice, this fact means that accommodations on ELP assessments generally address issues such as speech-language impairment or specific learning disabilities. Nonverbal students may need accommodations to provide alternate forms of expression for the speaking subtest, while EL students diagnosed with dyslexia may need special text presentations to ensure they can process written language appropriately. Setting accommodations, such as a quiet room or a small-group environment, may also be appropriate, though the effectiveness of such accommodations for ELs with disabilities is not yet well established (Rogers, Lazarus, & Thurlow, 2016). Accommodations like glossaries or simplified language are typically not appropriate on an ELP assessment.

Comparability between ELs with and without disabilities on ELP assessments would follow the same logic as for content assessments: when accommodations are used that have been shown to provide differential boost without affecting focal content or difficulty, scores may be aggregated. To date, however, we are not aware of any systematic reviews or meta-analyses of accommodations for ELs with disabilities specifically on ELP assessments. Whereas findings for some types of accommodations such as setting changes may be generalizable from non-ELs to ELs, this should not be assumed; more research is needed to establish the effectiveness and validity of accommodations for this population on this type of assessment.

### Recommendations

Given the various differences raised above, it is likely clear that ELP scores from different settings are rarely comparable. Certainly, scores from different ELP assessments are likely too different to be considered comparable. And, depending on the proposed use, even scores from the same ELP assessment in different settings may not be comparable in terms of how they should be used or interpreted, even if they are psychometrically equivalent. In light of this challenge, we offer the following recommendation for those who would wish to compare the ELP performance of ELs—as individuals or groups—across settings.

#### *Focus on Comparability of Uses and Interpretations, Rather Than Psychometric Comparability*

As discussed above, scores from the same ELP assessment (e.g., the WIDA ACCESS) may still “mean” different things in different settings because of how they are used. Specifically, though an ACCESS score of 500 can be assumed to be comparable across



all WIDA states from a psychometric perspective (i.e., the score represents the same level of achievement on the same construct), what that score means about a student's preparation and ability to participate in academic instruction without supports will still vary because of other state-specific contextual variables. A potential workaround to this challenge is to focus on how ELP scores relate to other relevant criteria, and frame comparability in terms of these relationships. Cook and colleagues (2012), for example, have argued for placing ELP proficiency cut scores at "the point at which EL students' academic content achievement ... becomes *less* related to their ELP" (p. 8), and provide three empirical methods for identifying such a point (Cook et al., 2012). As they show, this point may be in different places for different states, even if the states use the same ELP assessment and thus have scores that are psychometrically equivalent. Given the goals and uses for ELP scores, this approach to comparability may be useful for some policy and instructional decisions.

### *Reconsider How ELP Is Defined and Measured*

All of the ELP assessments currently in use take the same particular view of how language works. In the early days of NCLB, test developers and policy makers took a fairly narrow, literal reading of the law's language about "reading, writing, speaking, and comprehending" English and created the domain-based assessments described above. There are many scholars of second-language acquisition who would argue that this way of defining ELP—in terms of discrete domains, achievement standards, and a construct domain that is primarily composed of knowledge, rules, and skills—misunderstands, fundamentally, what language is or how it works (Larsen-Freeman, 2018). For those who take a more sociocognitive view of language, the idea that a learner could demonstrate their language skills in isolation using a written test (even a computer-based one) cuts directly against what language is and how it functions in the world. So, too, does the idea that one could isolate, define, and measure any one domain (e.g., "listening") in isolation from the others. The idea that learners progress toward proficiency in a linear fashion that one could trace for accountability—as opposed to cycling through various stages of development, some of which may look like regression in the near term—is also contrary to some thinkers' conceptualization of language as a construct (Larsen-Freeman, 2018; Pennycook, 2017; Valdés, 2018). While incorporating these views of language into ELP assessment might not necessarily improve test score comparability, it is possible that other ways of defining and measuring language might shift the conversation about how we understand EL students' achievement and language progress at both the aggregate and individual levels.

## **ESTIMATING GROWTH IN ACHIEVEMENT AND LANGUAGE PROFICIENCY FOR ENGLISH LEARNERS**

In several of the other chapters, test comparability is important because aggregate scores are being used to compare teachers and schools under various accountability regimens. For example, under ESSA, many states hold schools accountable for student growth, with the lowest-ranked schools subject to sanctions. While schools are also held accountable for student progress toward language proficiency, ELP assessment

scores are also used at the individual student level to help determine reclassification. Reclassification is a pivotal moment for ELs, representing the point at which the student is no longer formally considered an EL (Umansky & Reardon, 2014). Rigorous quasi-experimental research indicates that mistiming reclassification for ELs (either too early or too late) has major and often dire consequences for the educational attainment of students in this subgroup (Pompa & Villegas, 2017; Robinson, 2011; Umansky & Reardon, 2014).

While there are many factors that can influence the comparability of growth estimates based on ELP assessments, three bear additional discussion. First, under ESSA, educators and policy makers must set cut scores on ELP assessments associated with different language proficiency levels, which are often used in the reclassification determination (Robinson, 2011; Robinson-Cimpian & Thompson, 2015). Frequently, these cut scores are used to monitor a student's progress toward language proficiency (as we discuss below, progress toward proficiency is not the same as growth, or is at least not as operationalized). Second, the calibration and linking approaches used to develop the vertical scale can be consequential to inferences made about growth (Briggs & Weeks, 2009a). Third, particular approaches to estimating growth using ELP scores can also change inferences about students' language development (Matta & Soland, 2018). We briefly discuss the implications of test-based reclassification criteria for opportunity to learn before examining each of the three factors in turn.

### **Reclassification and Opportunity to Learn**

One of the primary reasons for monitoring ELs' English language development is to determine progress toward proficiency and, ultimately, reclassification. Research shows that the timing of when a student reclassifies can have major implications for ELs' opportunity to learn (Robinson, 2011; Robinson-Cimpian & Thompson, 2015; Thompson, 2015). This timing is based in large part on proficiency cut scores set on various ELP assessments. As discussed in the Introduction, EL status is intended to confer students with appropriate academic and language supports in the classroom and on assessments (Pompa & Villegas, 2017; Robinson, 2011; Umansky & Reardon, 2014). Once a student is reclassified, those supports diminish or disappear entirely. Thus, reclassifying a student too early can mean they lose needed supports, which, if actually appropriate and geared to their level of language development, is problematic. Reclassifying them too late, on the other hand, can potentially lead to stagnation (Dabach, 2014; Thompson, 2015) if they are grouped with students at much lower levels of language development and kept from grade-level subject-matter instruction.

Evidence of the importance of reclassification to students' opportunity to learn abounds. For example, Robinson (2011) showed that, in the district he studied, reclassification actually caused decreased reading test scores in later grades, suggesting that students had not received the instruction and preparation necessary from their teachers to forgo the supports associated with EL status. In related work, Robinson-Cimpian and Thompson (2015) showed that increasing the difficulty of attaining the test-based criteria for EL reclassification had significant positive effects on high school students' subsequent reading achievement (0.18 standard deviations) and graduation outcomes (11 percentage points). Recent research indicates that part of the reason reclassification

is consequential relates to the courses that ELs versus non-ELs can access, participate in, and benefit from. In particular, Umansky (2016) found evidence that certain course-assignment policies can lead to EL students being underrepresented in challenging courses or even shut out from them altogether. Taken together, these findings on reclassification indicate that calibrating associated test-based criteria is delicate, and that imperfections can have consequences for outcomes as fundamental as whether the student graduates.

There is also growing evidence that being a “long-term” EL can have negative consequences for students (Brooks, 2018). Though created to draw awareness to the needs of students who, according to assessments, have not acquired English in a “typical” time frame, the long-term EL label has acquired strongly negative connotations focused on students’ perceived deficits (Flores, Kleyn, & Menken, 2015; Thompson, 2015). Many of these negative connotations often arise because the failure of the educational system to prepare ELs (e.g., by restricting their access to core content [Umansky, 2016]) is wrongly interpreted as the student’s inability to learn English. Due to the misconceptions that often underline long-term EL status, evidence suggests that the long-term EL label can be stigmatizing for students, making them question their academic and language abilities (Dabach, 2014). In part as a result, students who take longer to exit EL status are less likely to graduate from high school or pursue postsecondary education (Heilig, 2011; Kanno & Cromley, 2013; Kao & Thompson, 2003).

Furthermore, some students end up being long-term ELs because they also have a learning disability, which can be conflated with not being proficient in English (Umansky, Thompson, & Díaz, 2017) and also complicates measurement of English language proficiency and implementation of policies based on related assessments. For example, there is evidence that existing policies and practices marginalize emergent bilinguals with disabilities by making reclassification improbable for students with intersecting disability and second-language acquisition needs (Schissel & Kangas, 2018). A technical challenge that can arise is that students with disabilities often do not have scores for all four English language proficiency subdomains, which makes producing the composite scores often used in reclassification difficult (Porter, Cook, & Sahakyan, 2019). From a policy standpoint, many states do not have exit criteria specific to students with disabilities, which means decisions are often left to districts (Thurlow, Shyyan, Lazarus, & Christensen, 2016). Altogether, these pieces of evidence suggest that ELs with disabilities are less likely to be reclassified, and that their higher likelihood of being long-term ELs can be as much a function of testing and reclassification policies as the language needs of the student.

In sum, students oftentimes end up being long-term ELs not because they lack English language proficiency, but because they have missed one of several test-based cut scores (Thompson, 2015). Thus, many of the deleterious effects of long-term EL status hinge on a single test-based criterion. These criteria often differ considerably across tests, states, and reclassification policies. In short, the cut scores used to reclassify students, which can be somewhat ad hoc across policy contexts, have major consequences for opportunity to learn.

### Setting ELP and Reading Cut Scores

In most states, reclassification is based on one or several test-based cut scores on both ELP and reading assessments (Linguanti & Cook, 2013). The comparability of these cut scores and, thereby, the meaning of the “fully English proficient” designation under reclassification is far from ensured. For example, states use different combinations of ELP scores in listening, speaking, reading, and writing subdomains, as well as overall ELP scores, to determine reclassification (Robinson-Cimpian & Thompson, 2015). Furthermore, the cut scores used on each assessment for the purposes of reclassification are not always consistent across states (in fact, even districts within some states can have different reclassification criteria) (Hill, Weston, & Hayes, 2014). Whereas one state may require that students attain a certain level of overall proficiency, as well as a certain level of proficiency for each subdomain (e.g., California), another might base reclassification on only the overall ELP cut score (e.g., Arizona). Given that the content of the various ELP assessments is not always the same (nor are the constructs necessarily defined in completely comparable ways), there is no guarantee that the meaning of designations like “proficient” are consistent across ELP assessments. In short, comparing proficiency designations and reclassification status across states and tests is often not justified.

### Construction of Vertical Scales

A primary way that proficiency cut scores on ELP assessments are used is to monitor students’ progress toward attaining proficiency over time, including under federal accountability. While having vertically scaled ELP assessments is not necessarily required for such purposes, the decisions about how vertical scales are constructed on ELP assessments (and tests more broadly) have implications for determining how quickly students are moving toward proficiency, especially in situations in which growth is being estimated. There are several broad approaches to developing vertical scales, but at least two are quite common: separate and concurrent calibration (Briggs & Weeks, 2009a; Tong & Kolen, 2007). Both typically assume that there are some common items administered to adjacent grades. Under separate calibration, parameters for items are estimated separately by grade but using the same item response theory (IRT) model. Item parameters from the shared items are then used to link the scales from adjacent grades post hoc. This linking process typically involves a linear transformation of the relevant item parameters. A disadvantage of this approach is that additional error is introduced because the linking parameters are estimated with error.

Under concurrent calibration, all item parameters for all grades are estimated in a single step. All the items from both grades are therefore calibrated to be on the same scale. However, if the construct is not consistent across grades, concurrent calibration can introduce bias across the scale that might be mitigated somewhat by cross-grade linking (Briggs & Weeks, 2009a). For a more thorough discussion of issues related to construct shift on ELP assessments with vertical scales, see Hansen and Monroe (2018).

On one hand, there is some evidence that such decisions often lead to fairly minimal differences in growth estimates. For example, Dadey and Briggs (2012) found that little of the considerable variability in the growth effect sizes across states on achievement tests could be explained by identifiable characteristics of the vertical scales. On the other hand, research indicates that the measure used can affect fundamental inferences about

growth, including across various IRT-based approaches (Briggs, 2013; Briggs & Weeks, 2009a; Seltzer, Frank, & Bryk, 1994). Even in the meta-analysis conducted by Dadey and Briggs (2012), there were substantial differences in growth estimates across states, just not that were due to clear IRT-based modeling decisions. Thus, one cannot be sure that gains on ELP assessments are comparable across contexts and tests.

### Estimating Growth

A primary reason many ELP assessments are vertically scaled is that stakeholders wish to use them to estimate growth in order to understand English language development trajectories. Several approaches to estimating growth have been employed, from complex models to other more rudimentary procedures. For example, under federal accountability, most states report progress rather than growth. That is, many states' ESSA plans require schools to report ELs' progress learning English, defined as whether students are moving up the proficiency designations on a test (e.g., going from "below basic" to "basic") over time. Using this coarse approach as a proxy for growth, the properties of the vertical scale may be less important than when comparing raw gains on that scale or fitting an actual growth model. Given differences in the ELP assessments used by states and how cut scores are set, state-by-state comparisons of progress toward language proficiency under federal law are not always justified.

In some cases, ELP assessments are actually used for estimation in a growth model. For example, several studies attempt to estimate time to reclassification, which is germane to evaluations of teachers, schools, and programs serving ELs (Burke et al., 2016; Loeb, Soland, & Fox, 2014; Umansky & Reardon, 2014). In particular, a range of studies examined the effects of different instructional environments on time to reclassification. Umansky and Reardon (2014) investigated time to reclassification among ELs in four linguistic instructional environments: English immersion, transitional bilingual, maintenance bilingual, and dual immersion. Similarly, Steele et al. (2017) used data from seven cohorts of language immersion lottery applicants to produce causal estimates of the effect of immersion on several outcomes, including time to reclassification. Both studies generally found positive effects associated with bilingual instruction. Beyond those two specific studies, one of the primary ways programs for ELs are evaluated is on ELP growth trends.

While such estimates of growth are likely sensitive to general modeling choices (e.g., whether to model growth as linear or quadratic), a bigger challenge for estimating growth for ELs is the instability of that subgroup.<sup>9</sup> Specifically, the ELs who evince the most growth are the most likely to be reclassified out of EL status, which means they no longer take ELP assessments and cannot be included in growth estimates. This

---

<sup>9</sup> Another major challenge in estimating time to reclassification is the fact that some students never experience reclassification, or at least do not experience it during the period for which data are available to create models. This issue of "censoring" is problematic because it will downwardly bias estimates of time to, or probability of, reclassification by excluding students for whom this event is not observed. In response to this known challenge, another increasingly used method for estimating time to reclassification is discrete-time survival analysis (Singer & Willett, 2003; Thompson, 2017), which supports the inclusion of ELs for whom reclassification has not yet occurred.



subgroup instability can be a major challenge to the inferences based on growth (Matta & Soland, 2019).

For example, when estimating time to reclassification, a problem arises: while a student's growth in ELP is an important predictor of when the student will no longer be considered an EL, the tests used to measure ELP are included directly in that determination. Because growth in ELP is a developmental process, employing the observed test scores in a time-to-reclassification model is therefore inappropriate because they are endogenous time-varying covariates measured with error (Kalbfleisch & Prentice, 2011). Thus, most studies examining predictors of time to reclassification do not account for how fast students' spoken and written proficiency in English is developing (Greenberg Motamedi, 2016; Kieffer & Parker, 2016; Matta & Soland, 2019). In plain terms, studies to understand what influences time to reclassification typically cannot include what is likely the most important predictor: a student's English language development.

Recently, Matta and Soland (2018) suggested a potential solution to the problem. They proposed a shared random effects model for the analysis of time to reclassification that accounts for English language development when reclassification decisions are conjunctive based on the total ELP score. Using their model, they improved predictions of time to reclassification by 17 percentage points relative to prior models (Matta & Soland, 2018). However, their model is technically complicated and is not simple to implement in standard statistical software (currently), which could limit its use among policy makers, educators, and applied researchers.

### Recommendations

Many of the recommendations we make regarding estimating growth for ELs are applicable to any attempt to use a test scale to understand developmental processes like language acquisition. Each recommendation parallels one of the factors we mention above that can affect score (and growth estimate) comparability: setting cut scores, developing vertical scales, and estimating growth.

#### *Set Cut Scores Carefully, Allow for Flexibility*

Under current federal law, states must use test-based cut scores of some kind to determine when ELs have reached proficiency. With that stricture in mind, there are several steps that can be taken to help safeguard against unintended consequences like students lingering unnecessarily in EL status. First, cut scores can be set thoughtfully based on the construct, test scale, and meaningful criteria identified by language and teaching experts. Relatedly, such cut scores should factor in measurement error in the scores. Second, layering many test-based cut scores may increase the likelihood that students remain ELs not because they need more language instruction to be successful in English-based classrooms, but because they have missed a single cutoff by chance. Thus, policy makers may wish to avoid layering too many test-based cut scores when making reclassification decisions. Third, local educators and parents should be given some latitude to reclassify students who, based on professional judgment, are ready for an English-based classroom even if falling short of a test-based criterion. Finally, rigorous methods like those used by Robinson (2011) should be used to calibrate cut



scores to help ensure that reclassified students are sufficiently prepared for the content without lingering in EL status, especially given the implications of reclassification for course access.

***Develop Vertical Scales According to Best Practices and Be Clear About the Limitations of Such Scales***

Fortunately, most of the ELP assessments we examined for this chapter use a vertical scale calibrated based on one of the two IRT-based approaches we describe (or a highly related method). States that are interested in understanding ELP growth trajectories should use such a scale, even if a vertical scale is not required to monitor progress under federal accountability. Furthermore, measurement experts should be clear about the limitations of the scales they have constructed, including what grade bands can appropriately be combined in a single growth model and whether raw scale score gains can and should be compared.

***Move Toward Growth Rather Than Progress***

Under federal law, many states use a crude proxy for growth in ELP. Specifically, they examine progress defined as the percentage of students moving from one ELP proficiency designation to another. Research shows that growth and gaps in growth based on proficiency bands can be highly sensitive to the cut scores used (Ho, 2008). Thus, policy makers may benefit from comparing schools' progress to actual estimates of school-level growth on ELP assessments. For the latter, growth models should be constructed in a way that accounts for the shifting nature of the EL subgroup, especially when being used to evaluate the effectiveness of educational programs and settings.

## CONCLUSION

In this chapter, we have sought to identify and explain several comparability issues that arise for comparisons involving students identified as English learners. Importantly, a foundational issue pertains to the population itself: for test-based reasons, even the EL label is not comparable from one setting to the next. Beyond this issue (as one might still want or need to make comparisons involving EL students), we address three other areas in particular: (1) comparisons of ELs to one another and to non-ELs within a state on assessments of academic content, (2) comparisons of ELs to one another across states on assessments of ELP, and (3) comparisons of linguistic growth and development for ELs across states.

Across all these scenarios, several comparability challenges are presented, none of which can be eliminated entirely (at least at present). However, we have sought to provide recommendations for considerations and checks that test users and developers may employ, both to minimize comparability issues where this is possible and, where it is not, to quantify the extent to which scores lack comparability. Many of these checks take the form of post hoc analyses of assessment data to attempt to quantify things like whether accommodations have altered the construct of an academic content assessment or whether two ELP scores mean the same thing about a student's preparedness to access grade-level

instruction in English. As this approach suggests, evidence of comparability should be sought and collected as part of an ongoing validation effort for any and all score uses and interpretations. Concurrent to such efforts, comparisons should be made carefully, always keeping EL students' strengths and vulnerabilities at the forefront when deciding how scores will be used, interpreted, and compared.

## REFERENCES

- Abedi, J. (2004). The No Child Left Behind Act and English language learners: Assessment and accountability issues. *Educational Researcher*, 33(1), 4–14.
- Abedi, J. (2006). Psychometric issues in the ELL assessment and special education eligibility. *The Teachers College Record*, 108(11), 2282–2303.
- Abedi, J. (Ed.). (2007). *English language proficiency assessment in the nation: Current status and future practice* (pp. 3–10). Davis, CA: University of California, Davis, School of Education.
- Abedi, J. (2009). Computer testing as a form of accommodation for English language learners. *Educational Assessment*, 14(3–4), 195–211. <https://doi.org/10.1080/10627190903448851>.
- Abedi, J. (2016). Utilizing accommodations in assessment. In E. Shohamy, I. G. Or, & S. May (Eds.), *Language testing and assessment: Encyclopedia of language and education* (3rd ed., pp. 1–20). Cham, Switzerland: Springer. [https://doi.org/10.1007/978-3-319-02326-7\\_21-1](https://doi.org/10.1007/978-3-319-02326-7_21-1).
- Abedi, J., & Ewers, N. (2013). *Accommodations for English language learners and students with disabilities: A research-based decision algorithm*. Smarter Balanced Assessment Consortium.
- Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education*, 14(3), 219–234.
- AERA (American Educational Research Association), APA (American Psychological Association), & NCME (National Council on Measurement in Education). (2014). *Standards for educational and psychological testing*. Washington, DC: APA.
- Aguirre-Muñoz, Z., & Boscardin, C. K. (2008). Opportunity to learn and English learner achievement: Is increased content exposure beneficial? *Journal of Latinos and Education*, 7(3), 186–205.
- Allard, E. C. (2016). Latecomers: The sources and impacts of late arrival among adolescent immigrant students. *Anthropology & Education Quarterly*, 47(4), 366–384.
- Avenia-Tapper, B., & Llosa, L. (2015). Construct relevant or irrelevant?: The role of linguistic complexity in the assessment of English language learners' science knowledge. *Educational Assessment*, 20(2), 95–111. <https://doi.org/10.1080/10627197.2015.1028622>.
- Bailey, A. L., Butler, F., & Sato, E. (2007). Standards-to-standards linkage under Title III: Exploring common language demands in ELD and science standards. *Applied Measurement in Education*, 20(1), 53–78.
- Briggs, D. C. (2013). Measuring growth with vertical scales. *Journal of Educational Measurement*, 50(2), 204–226. <https://doi.org/10.1111/jedm.12011>.
- Briggs, D. C., & Weeks, J. P. (2009a). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues and Practice*, 28(4), 3–14.
- Briggs, D. C., & Weeks, J. P. (2009b). The sensitivity of value-added modeling to the creation of a vertical score scale. *Education*, 4(4), 384–414.
- Brooks, M. D. (2018). Pushing past myths: Designing instruction for long term English learners. *TESOL Quarterly*, 52(1), 221–233.
- Burke, A. M., Morita-Mullaney, T., & Singh, M. (2016). Indiana emergent bilingual student time to reclassification: A survival analysis. *American Educational Research Journal*, 53(5), 1310–1342.
- Cai, L., & Hansen, M. (2018). Improving educational assessment: Multivariate statistical methods. *Policy Insights from the Behavioral and Brain Sciences*, 5(1), 19–24. <https://doi.org/10.1177/2372732217747006>.
- California Department of Education. (2012, November). *Overview of the California English language development standards and proficiency level descriptors*. Retrieved from <http://www.cde.ca.gov/sp/el/er/documents/eldstndpublication14.pdf>.
- Carlson, D., & Knowles, J. E. (2016). The effect of English language learner reclassification on student ACT scores, high school graduation, and postsecondary enrollment: Regression discontinuity evidence from Wisconsin. *Journal of Policy Analysis and Management*, 35(3), 559–586. <https://doi.org/10.1002/pam.21908>.

- Carroll, P. E. (2012). *Examining the validity of classifications from an English language proficiency assessment for English language learners and native English speakers in fifth grade*. Retrieved from <http://escholarship.org/uc/item/2332d3j5>.
- Carroll, P. E., & Bailey, A. L. (2015). Do decision rules matter? A descriptive study of English language proficiency assessment classifications for English-language learners and native English speakers in fifth grade. *Language Testing*, 1–30. <https://doi.org/10.1177/0265532215576380>.
- CCSSO (Council of Chief State School Officers). (2013). *English language proficiency (ELP) standards with correspondences to K-12 English language arts (ELA), mathematics, and science practices, K-12 ELA standards, and 6-12 literacy standards*. Retrieved from [http://www.elpa21.org/sites/default/files/Final%204\\_30%20ELPA21%20Standards\\_1.pdf](http://www.elpa21.org/sites/default/files/Final%204_30%20ELPA21%20Standards_1.pdf).
- CCSSO. (2016). *Major provisions of Every Student Succeeds Act (ESSA) related to the education of English learners*. Washington, DC: CCSSO.
- Cook, H. G., Boals, T., & Lundberg, T. (2011). Academic achievement for English learners: What can we reasonably expect? *Phi Delta Kappan*, 93(3), 66–69.
- Cook, H. G., Linquanti, R., Chinen, M., & Jung, H. (2012). *National evaluation of Title III implementation supplemental report: Exploring approaches to setting English language proficiency performance criteria and monitoring English learner progress*. Washington, DC: Office of Planning, Evaluation and Policy Development, U.S. Department of Education.
- Dabach, D. B. (2014). “I am not a shelter!”: Stigma and social boundaries in teachers’ accounts of students’ experience in separate “sheltered” English learner classrooms. *Journal of Education for Students Placed at Risk*, 19(2), 98–124.
- Dadey, N., & Briggs, D. C. (2012). A Meta-Analysis of Growth Trends from Vertically Scaled Assessments. *Practical Assessment, Research & Evaluation*, 17.
- Faulkner-Bond, M., & Forte, E. (2016). English learners and accountability: The promise, pitfalls, and peculiarity of assessing language minorities via large-scale assessment. In C. Wells & M. Faulkner-Bond (Eds.), *Educational measurement: From foundations to future* (pp. 395–415). New York: Guilford Press.
- Faulkner-Bond, M., Wolf, M. K., Wells, C. S., & Sireci, S. G. (2018). Exploring the factor structure of a K–12 English language proficiency assessment. *Language Assessment Quarterly*, 1–20. <https://doi.org/10.1080/15434303.2017.1419247>.
- Flores, N., Kleyn, T., & Menken, K. (2015). Looking holistically in a climate of partiality: Identities of students labeled long-term English language learners. *Journal of Language, Identity & Education*, 14(2), 113–132.
- Frantz, R. S., Bailey, A. L., Starr, L., & Perea, L. (2014). Measuring academic language proficiency in school-age English language proficiency assessments under new college and career readiness standards in the United States. *Language Assessment Quarterly*, 11(4), 432–457. <https://doi.org/10.1080/15434303.2014.959123>.
- Gándara, P., & Santibañez, L. (2016). The teachers our English language learners need. *Educational Leadership*, 73(5), 32–37.
- Greenberg Motamedi, J. (2015). *Time to reclassification: How long does it take English learner students in Washington road map districts to develop English proficiency?* REL 2015-092. Regional Educational Laboratory Northwest. Retrieved from <https://eric.ed.gov/?id=ED558159>.
- Hansen, M., & Monroe, S. (2018). Linking not-quite-vertical scales through multidimensional item response theory. *Measurement: Interdisciplinary Research and Perspectives*, 16(3), 155–167.
- Heilig, J. V. (2011). Understanding the interaction between high-stakes graduation tests and English learners. *Teachers College Record*, 113(12), 2633–2669.
- Hill, L. E., Weston, M., & Hayes, J. M. (2014). *Reclassification of English learner students in California*. San Francisco, CA: Public Policy Institute of California.
- Ho, A. D. (2008). The problem with “proficiency”: Limitations of statistics and policy under No Child Left Behind. *Educational Researcher*, 37(6), 351–360.
- Kachchaf, R., Noble, T., Rosebery, A., O’Connor, C., Warren, B., & Wang, Y. (2016). A closer look at linguistic complexity: Pinpointing individual linguistic features of science multiple-choice items associated with English language learner performance. *Bilingual Research Journal*, 39(2), 152–166. <https://doi.org/10.1080/15235882.2016.1169455>.

- Kalbfleisch, J. D., & Prentice, R. L. (2011). *The statistical analysis of failure time data* (Vol. 360). Hoboken, NJ: John Wiley & Sons.
- Kanno, Y., & Cromley, J. G. (2013). English language learners' access to and attainment in postsecondary education. *Tesol Quarterly*, 47(1), 89–121.
- Kao, G., & Thompson, J. S. (2003). Racial and ethnic stratification in educational achievement and attainment. *Annual review of sociology*, 29(1), 417–442.
- Kieffer, M. J. & Parker, C. E. (2016). *Patterns of English learner student reclassification in New York City public schools* (REL 2017–200). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast & Islands. Retrieved from <http://ies.ed.gov/ncee/edlabs>.
- Kieffer, M. J., Rivera, M., & Francis, D. J. (2012). *Practical guidelines for the education of English language learners: Research-based recommendations for the use of accommodations in large-scale assessments*. 2012 update. Portsmouth, NH: RMC Research Corporation, Center on Instruction.
- Kopriva, R. J. (2014). Second-generation challenges for making content assessments accessible for ELLs. *Applied Measurement in Education*, 27(4), 301–306. <https://doi.org/10.1080/08957347.2014.944311>.
- Kopriva, R. J., & Wright, L. J. (2017). Score processes in assessing academic content of non-native speakers: Literature review and ONPAR summary. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments: The use of response processes* (pp. 100–112). New York: Routledge.
- Lane, S., & Leventhal, B. (2015). Psychometric challenges in assessing English language learners and students with disabilities. *Review of Research in Education*, 39(1), 165–214. <https://doi.org/10.3102/0091732X14556073>.
- Larsen-Freeman, D. (2018). Looking ahead: Future directions in, and future research into, second language acquisition. *Foreign Language Annals*, 51(1), 55–72. <https://doi.org/10.1111/flan.12314>.
- Li, H., & Suen, H. K. (2012). The effects of test accommodations for English language learners: A meta-analysis. *Applied Measurement in Education*, 25(4), 327–346. <https://doi.org/10.1080/08957347.2012.714690>.
- Linquanti, R., & Cook, H. G. (2013). *Toward a common definition of English learner: Guidance for states and state assessment consortia in defining and addressing policy and technical issues and options*. Washington, DC: Council of Chief State School Officers.
- Liu, K. K., Ward, J. M., Thurlow, M. L., & Christensen, L. L. (2015). Large-scale assessment and English language learners with disabilities. *Educational Policy*, 31(5), 551–583. <https://doi.org/10.1177/0895904815613443>.
- Loeb, S., Soland, J., & Fox, L. (2014). Is a good teacher a good teacher for all? Comparing value-added of teachers with their English learners and non-English learners. *Educational Evaluation and Policy Analysis*, 36(4), 457–475.
- Logan-Terry, A., & Wright, L. J. (2010). Making thinking visible: An analysis of English language learners' interactions with access-based science assessment items. *AccELLerate!*, 2(4), 11–14.
- Martiniello, M. (2008). Language and the performance of English-language learners in math word problems. *Harvard Educational Review*, 78(2), 333–368.
- Martiniello, M. (2009). Linguistic complexity, schematic representations, and differential item functioning for English language learners in math tests. *Educational Assessment*, 14(3–4), 160–179. <https://doi.org/10.1080/10627190903422906>.
- Matta, T. H., & Soland, J. (2019). Predicting time to reclassification for English learners: A joint modeling approach. *Journal of Educational and Behavioral Statistics*, 44(1), 78–102. <https://doi.org/10.3102/1076998618791259>.
- NASEM (National Academies of Sciences, Engineering, and Medicine). (2017). *Promoting the educational success of children and youth learning English: Promising futures*. Washington, DC: The National Academies Press.
- Noble, T., Rosebery, A., Kachchaf, R., & Suarez, C. (2016). *A handbook for improving the validity of multiple-choice science test items for English language learners*. Cambridge, MA: TERC.
- Noble, T., Rosebery, A., Suarez, C., Warren, B., & O'Connor, M. C. (2014). Science assessments and English language learners: Validity evidence based on response processes. *Applied Measurement in Education*, 27(4), 248–260. <https://doi.org/10.1080/08957347.2014.944309>.



- PARCC (Partnership for Assessment of Readiness for College and Careers). (2017, August 3). *PARCC accessibility features and accommodations manual: Guidance for districts and decision-making teams to ensure that PARCC summative assessments produce valid results for all students* (6th ed.). Available from <http://avocet.pearson.com/PARCC/Home>.
- Pennock-Roman, M., & Rivera, C. (2011). Mean effects of test accommodations for ELLs and non-ELLs: A meta-analysis of experimental studies. *Educational Measurement: Issues and Practice*, 30(3), 10–28. <https://doi.org/10.1111/j.1745-3992.2011.00207.x>.
- Pennycook, A. (2017). *Posthumanist applied linguistics*. New York: Routledge.
- Pompa, D., & Villegas, L. (2017). *Analyzing state ESSA plans for English learner accountability: A framework for community stakeholders*. Migration Policy Institute.
- Porter, T., Cook, H. G., & Sahakyan, N. (2019). Less than four domains: Creating an overall composite score for English learners with Individualized Education Plans. Retrieved from <https://wida.wisc.edu/sites/default/files/resource/Less-Than-Four-Domains.pdf>.
- Pottinger, J. S. (1970, May 25). *Identification of discrimination and denial of services on the basis of national origin*. Retrieved from <http://www2.ed.gov/about/offices/list/ocr/docs/lau1970.html>.
- Robinson, J. P. (2011). Evaluating criteria for English learner reclassification: A causal-effects approach using a binding-score regression discontinuity design with instrumental variables. *Educational Evaluation and Policy Analysis*, 33(3), 267–292.
- Robinson-Cimpian, J. P., & Thompson, K. D. (2015). The effects of changing test-based policies for reclassifying English learners. *Journal of Policy Analysis and Management*, 35(2), 279–305. <http://doi.org/10.1002/pam.21882>.
- Rogers, C. M., Lazarus, S., & Thurlow, M. L. (2016). *A summary of the research on the effects of test accommodations: 2013–2014* (NCEO Report No. 402). Retrieved from <https://nceo.umn.edu/docs/OnlinePubs/Report402/NCEORep402.pdf>.
- Schissel, J. L., & Kangas, S. E. (2018). Reclassification of emergent bilinguals with disabilities: The intersectionality of improbabilities. *Language Policy*, 17(4), 567–589.
- Seltzer, M. H., Frank, K. A., & Bryk, A. S. (1994). The metric matters: The sensitivity of conclusions about growth in student achievement to choice of metric. *Educational Evaluation and Policy Analysis*, 16(1), 41–49.
- Shyyan, V., Thurlow, M., Lazarus, S., Christensen, L., Corpe, J., Rogers, C., & Larson, E. (2017). *Data informed accessibility: A review of the literature* (p. 42). Retrieved from <https://nceo.umn.edu/docs/OnlinePubs/DIAMONDLiteratureReviewReport.pdf>.
- Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford, UK: Oxford University Press.
- Sireci, S. G., & Faulkner-Bond, M. (2015). Promoting validity in the assessment of English learners. *Review of Research in Education*, 39(1), 215–252. <https://doi.org/10.3102/0091732X14557003>.
- Sireci, S. G., Scarpatti, S. E., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research*, 75(4), 457–490. <https://doi.org/10.3102/00346543075004457>.
- Sireci, S. G., & Wells, C. S. (2010). Evaluating the comparability of English and Spanish video accommodations for English language learners. In P. C. Winter (Ed.), *Evaluating the comparability of scores from achievement test variations* (pp. 33–68). Washington, DC: Council of Chief State School Officers.
- Smarter Balanced Assessment Consortium. (2014, January 26). *Smarter Balanced Assessment Consortium: Accessibility and accommodations framework*. Retrieved from <https://portal.smarterbalanced.org/library/en/accessibility-and-accommodations-framework.pdf>.
- Smith, W. L. (1990, April 6). *OCR policy regarding the treatment of national origin minority students who are limited English proficient*. Retrieved from [https://www2.ed.gov/about/offices/list/ocr/docs/lau1990\\_and\\_1985.html](https://www2.ed.gov/about/offices/list/ocr/docs/lau1990_and_1985.html).
- Solano-Flores, G. (2012). *Translation accommodations framework for testing English language learners in mathematics*. Smarter Balanced Assessment Consortium.
- Solano-Flores, G. (2014). Probabilistic approaches to examining linguistic features of test items and their effect on the performance of English language learners. *Applied Measurement in Education*, 27(4), 236–247. <https://doi.org/10.1080/08957347.2014.944308>.

- Solano-Flores, G., & Li, M. (2009). Language variation and score variation in the testing of English language learners, native Spanish speakers. *Educational Assessment*, 14(3–4), 180–194. <https://doi.org/10.1080/10627190903422880>.
- Solano-Flores, G., Wang, C., Kachchaf, R., Soltero-Gonzalez, L., & Nguyen-Le, K. (2014). Developing testing accommodations for English language learners: Illustrations as visual supports for item accessibility. *Educational Assessment*, 19(4), 267–283. <https://doi.org/10.1080/10627197.2014.964116>.
- Steele, J. L., Slater, R. O., Zammaro, G., Miller, T., Li, J., Burkhauser, S., & Bacon, M. (2017). Effects of dual-language immersion programs on student achievement: Evidence from lottery data. *American Educational Research Journal*, 54(1), 282S–306S.
- Thompson, K. D. (2015). Questioning the long-term English learner label: How categorization can blind us to students' abilities. *Teachers College Record*, 117(12).
- Thompson, K. D. (2017). English learners' time to reclassification: An analysis. *Educational Policy*, 31(3), 330–363. <https://doi.org/10.1177/0895904815598394>.
- Thurlow, M. L., & Kopriya, R. J. (2015). Advancing accessibility and accommodations in content assessments for students with disabilities and English learners. *Review of Research in Education*, 39(1), 331–369. <https://doi.org/10.3102/0091732X14556076>.
- Thurlow, M. L., Shyyan, V. V., Lazarus, S. S., & Christensen, L. L. (2016). *Providing English language development services to English learners with disabilities: Approaches to making exit decisions*. NCEO Report 404. National Center on Educational Outcomes.
- Tong, Y., & Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education*, 20(2), 227–253.
- Umansky, I. M. (2016). Leveled and exclusionary tracking: English learners' access to academic content in middle school. *American Educational Research Journal*, 53(6), 1792–1833. <https://doi.org/10.3102/0002831216675404>.
- Umansky, I. M., & Reardon, S. F. (2014). Reclassification patterns among Latino English learner students in bilingual, dual immersion, and English immersion classrooms. *American Educational Research Journal*, 51(5), 879–912.
- Umansky, I. M., Thompson, K. D., & Díaz, G. (2017). Using an ever-English learner framework to examine disproportionality in special education. *Exceptional Children*, 84(1), 76–96. <https://doi.org/10.1177/0014402917707470>.
- Valdés, G. (2018). Analyzing the curricularization of language in two-way immersion education: Restating two cautionary notes. *Bilingual Research Journal*, 41(4), 388–412.
- WIDA Consortium. (2012). *2012 amplification of the English language development standards, kindergarten–grade 12*. Board of Regents of the University of Wisconsin System.
- Williams, M. L. (1991, September 27). *Policy update on schools' obligations toward national origin minority students with limited-English proficiency (LEP students)*. Retrieved from <https://www2.ed.gov/about/offices/list/ocr/docs/lau1991.html>.
- Wolf, M. K., & Leon, S. (2009). An investigation of the language demands in content assessments for English language learners. *Educational Assessment*, 14(3–4), 139–159. <https://doi.org/10.1080/10627190903425883>.
- Working Group on ELL Policy. (2010). *Improving educational outcomes for English language learners: Recommendations for the reauthorization of the Elementary and Secondary Education Act*. Palo Alto, CA: Authors. Retrieved from <http://ellpolicy.org/esea>.
- Wu, Y.-C., Liu, K., Thurlow, M. L., & Albus, D. (2019). *Number and percent of English learners with disabilities by disability category for 2013-14 to 2016-17* (Data Analytics No. 9). Retrieved from [https://tableau.ahc.umn.edu/t/ICI/views/ELswithDisabilitiesfinalDataAnalytic9/Story1?iframeSizedToWindow=true&:embed=y&:showAppBanner=false&:display\\_count=no&:showVizHome=no#1https://wested.onelogin.com/login](https://tableau.ahc.umn.edu/t/ICI/views/ELswithDisabilitiesfinalDataAnalytic9/Story1?iframeSizedToWindow=true&:embed=y&:showAppBanner=false&:display_count=no&:showVizHome=no#1https://wested.onelogin.com/login).





# Comparability When Assessing Individuals with Disabilities

Stephen G. Sireci and Maura O’Riordan, *University of Massachusetts Amherst*

## CONTENTS

INTRODUCTION .....	178
For Whom Are Test Accommodations Intended? .....	179
Defining Testing Purposes and Constructs Measured .....	180
VALIDITY AND ACCOMMODATIONS .....	181
Accommodations Versus Designated Supports .....	182
Accommodations Versus Modifications .....	183
Validity and Fairness .....	183
TYPES OF TEST ACCOMMODATIONS .....	184
Accommodations on College Admissions Tests .....	186
COMPARABILITY ISSUES IN TEST ACCOMMODATIONS .....	188
Evaluating Comparability .....	189
REVIEWING THE LITERATURE ON EVALUATING TEST ACCOMMODATIONS .....	190
Timing Accommodations .....	191
Read Aloud as a Presentation Accommodation .....	192
Response Accommodations .....	193
Setting Accommodations .....	194
Equipment and Devices Accommodations .....	194
Summary of the Review .....	195
Deriving Recommendations from the Literature .....	195
ALTERNATE ASSESSMENTS FOR STUDENTS WITH SEVERE DISABILITIES .....	196
Research on Alternate Assessments .....	199
TEST ACCOMMODATIONS AND TECHNOLOGY .....	199
CONCLUSION .....	200
REFERENCES .....	201

## INTRODUCTION

In this chapter, we discuss how deviations from typical test administration conditions affect the comparability of scores from educational tests. This discussion requires the definition of three terms: comparability, standardized, and accommodation. In its most general sense, *comparability* refers to the degree to which examinees' scores on a test can be meaningfully compared. To facilitate comparable scores across examinees, most educational tests are *standardized*, which means the test content, administration conditions, and scoring procedures are the same (uniform) for all test takers. Thus, standardization is designed to promote fairness in testing by providing a level playing field for all examinees. However, just as stairs make it difficult for people using wheelchairs to enter a building, features of a standardized testing situation may make it difficult for individuals with disabilities to fully interact with the assessment process. In fact, some features of a standardized testing system may prevent individuals with disabilities from demonstrating their knowledge, skills, and abilities. For this reason, testing agencies often provide accommodations to the standardized testing situation. *Accommodations* refer to changes in (1) the presentation of test content, (2) the setting in which a test is administered, (3) the manner in which examinees provide responses to test questions, (4) the amount of time given to examinees to complete a test or sections of a test, and (5) the use of additional resources or devices on the test. There are nuances related to the term "accommodation" that are explained later in this chapter. The point to bear in mind is that accommodations to standardized tests are oxymoronic (Sireci, 2005), in that accommodations change standardized procedures, which are designed to be uniform. Thus, changes to standard testing conditions may threaten score comparability across tests taken under standard and nonstandard conditions. However, without such changes, many examinees with disabilities could not be properly assessed. We confront this dilemma in this chapter.

The chapter focuses on issues and practices related to test accommodations for students with disabilities (SWD) and how they relate to issues of score comparability and fairness in assessment. We begin with a description of the types of students for whom accommodations are intended and then describe some relevant psychometric concepts in this area, such as validity, construct, construct representation, and construct-irrelevant variance. Test development procedures designed to make tests more accessible to SWD, such as universal test design, are also covered. We then describe current practices in providing test accommodations on educational tests. The issue of *flagging* test scores (i.e., providing a demarcation on a test score report that the test was taken with an accommodation) is also discussed. We also present a brief review of the literature on the effects of test accommodation on score interpretation and score comparability, and we end with suggestions for future research and practice aimed to facilitate comparability of test scores across individuals with and without disabilities.

### For Whom Are Test Accommodations Intended?

There are generally four groups of examinees most often considered for accommodations: (1) individuals with disabilities, (2) English learners (ELs),<sup>1</sup> (3) ELs with disabilities, and (4) individuals with severe cognitive or physical impairments. Each of these four groups varies widely in profile. Thus, it is important to note that two students with the same group status may learn differently and have different needs, and so there is no one-size-fits-all accommodation for any group.

#### *Individuals with Disabilities*

In the United States, the Individuals with Disabilities Education Act of 2004 (IDEA) (20 U.S.C. § 1400) was designed to ensure equity and accountability in education for children with disabilities. To demonstrate the diversity within the group label *individuals with disabilities* (hereafter referred to as “students with disabilities” or SWD), IDEA identified 13 types of students who may qualify for accommodations due to disability. They are students with autism, deaf-blindness, deafness, emotional disturbance, hearing impairment, intellectual disability, multiple disabilities, orthopedic impairment, other health impairment, specific learning disability, speech or language impairment, traumatic brain injury, and visual impairment. Even within these subgroups, disabilities can vary widely. For example, “other health impairment” refers to a person with “limited strength, vitality, or alertness” (IDEA, 2004, 34 C.F.R. 300.8(c)(9)) and can vary from asthma, to attention deficit hyperactivity disorder, to epilepsy or diabetes. From an assessment standpoint, none of those students would benefit from the same accommodation.

Within schools, two types of plans exist for SWD who are having difficulty accessing grade-level material due to a disability. The first is an individualized education plan (IEP). An IEP is created for an individual student with a disability by a group of people that includes parents, special educators, and teachers. IEPs are mandated by IDEA (IDEA, 2004, 34 C.F.R. 300.32) and must include the student’s present level of academic achievement, measurable annual goals, a description of how progress toward those goals will be measured, a statement of the services that will be provided to the student, a description of the extent to which the student will not participate with students in the class who do not have a disability, and a date the services will start along with the anticipated schedule for the outlined plan. IEPs are designed for students who fall under one of the 13 disability categories listed in IDEA, and the annual IEP meeting and subsequent plan is necessarily extensive to meet the outlined requirements.

The second type of plan SWD may receive is a 504 plan. This plan refers to section 504 of the Rehabilitation Act of 1973 (34 C.F.R. Part 104.4), which is a broad civil rights law protecting all people with disabilities. Section 504 mandates that individuals with disabilities have equal rights to fully participate in programs that receive federal

---

<sup>1</sup> We use the term English learners here because that is the most common group of students accommodated in the United States due to limited proficiency in the language in which the test is administered. However, limited language proficiency generalizes to any situation where an examinee is tested in a language in which they are not fully proficient (see ITC, 2018, and Chapter 6, Comparability When Assessing English Learner Students).

funding. This applies to schools for SWD who do not require IEPs, such as students with physical or mental impairments that limit their daily activities. 504 plans can be as comprehensive as IEPs, but by law it is only mandated to note the description of services provided to ensure equal participation. Thus, typically 504 plans are not as comprehensive as IEPs.

### ***English Learners and English Learners with Disabilities***

Within-group heterogeneity is not unique to SWD; ELs are just as diverse. More than 400 different languages were reported to be spoken by ELs in the United States in the 2015–2016 school year (DOEd, 2016). In addition to the many languages spoken, students are at varying levels of English language acquisition. To make matters even more complicated, some ELs may have a disability, which may require even more specialized accommodations.

### ***Individuals with More Severe Disabilities***

The final group encompasses individuals with severe cognitive or physical impairments. This group of students is typically not able to access the assessment even with accommodations, instead participating in *alternate assessments*. In statewide assessment systems in the United States, alternate assessments typically have different (modified) achievement standards than those measured by the general assessments (NCEO, 2016).

## **Defining Testing Purposes and Constructs Measured**

All tests are developed to serve one or more purposes. For example, a state department of education may develop an eighth grade mathematics assessment for the purpose of measuring eighth grade students' mastery of a statewide eighth grade mathematics curriculum, and informing parents and educators of that mastery or lack thereof (see Fremer & Wall [2004] or Sireci & Gandara [2016] for more examples of purposes of educational tests). After defining the purpose for which a test is to be created, a testing program must define the "construct" to be measured by the test. Cronbach and Meehl (1955) introduced the term *construct* to describe "some postulated attribute of people, assumed to be reflected in test performance" (p. 283). Essentially, the knowledge, skill, or other attribute measured by a test is a construct.

The constructs measured by educational and psychological tests cannot be directly seen, and so they are often thought of as "underlying" constructs or hypothetical traits. The *Standards for Educational and Psychological Testing* (hereafter referred to as the *Standards*) define a construct as "the concept or characteristic the test is intended to measure" (AERA, APA, & NCME, 2014, p. 11). For example, the aforementioned eighth grade mathematics test may define the construct measured using the statewide curriculum framework for eighth grade mathematics. As another example, proficiency in occupational therapy may be the construct targeted by an occupational therapy licensure test. In that situation, the construct may be defined by an analysis of the tasks carried out by occupational therapists and the knowledge and skills they need to successfully complete those tasks.

Defining the constructs to be measured by a test is critically important to understand the characteristics of examinees that are reflected in test scores. In addition, the construct definition helps evaluate whether a specific accommodation may alter the constructs measured (i.e., change what the test measures). For example, if quality of handwriting is considered part of the construct measured by a writing test, allowing an examinee to dictate their responses to a writing prompt to a scribe may change the construct from writing proficiency to speaking proficiency. For this reason, the AERA et al. (2014) *Standards* state, “The test developer should set forth clearly how test scores are intended to be interpreted and consequently used ... and the construct or constructs that the test is intended to assess should be described clearly” (p. 23).

Defining the purposes of a testing program and the constructs measured by the test are not just the first two steps in developing a test; they also set the stage for evaluating the validity, utility, and fairness of a testing program. As part of that evaluation, the degree to which various accommodations are appropriate, and the degree to which they lead to comparable or noncomparable scores, must be addressed. These issues are more broadly characterized within the psychometric concept of *validity*, which we turn to next.

## VALIDITY AND ACCOMMODATIONS

The AERA et al. (2014) *Standards* define validity as “the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests” (p. 11). Thus, with respect to educational testing, the question of validity is not “Is this test valid?” but rather “Are the interpretations based on the test scores valid for this particular use?” This distinction might seem trivial, but it is important to note that it takes the concept of validity away from the idea that it is a property of a test and instead grounds it in very specific contexts based on how scores are interpreted and used.

A topic related to validity is “validation,” which is a process through which evidence is accumulated to support specific score interpretations of specific proposed uses of tests. The idea of validating the interpretation and uses of a test, rather than validating a test itself, is fundamental to understanding 21st-century notions of validity. Validating test use creates the need for test developers to clearly define the purpose of an assessment; it ties the interpretations of scores strictly to that proposed use.

The AERA et al. (2014) *Standards* describe five sources of validity evidence that can be used to evaluate the use of a test for a particular purpose. These sources are validity evidence based on test content, response processes, internal structure, relations to other variables, and consequences of testing. The first four of these sources can be used to evaluate how well the test measures its intended construct, and so they can likewise be used to evaluate the degree to which accommodations may alter the construct measured (Sireci, Banda, & Wells, 2018; Sireci & Faulkner-Bond, 2015). Thus, in addition to confirming that a test generally measures its intended construct, validation also requires confirmation that the construct is measured with similar quality for all test takers.

Samuel Messick, one of the most prolific and respected validity theorists of all time, claimed that problems in fair and valid assessment arise from either construct underrepresentation or construct-irrelevant variance. As he put it, “Tests are imperfect



measures of constructs because they either leave out something that should be included ... or else include something that should be left out, or both" (Messick, 1989, p. 34).

*Construct underrepresentation* refers to a situation in which a test measures only a portion of the intended construct and leaves important knowledge, skills, and abilities untested. For example, if Spanish proficiency were operationally defined as reading, writing, speaking, and listening in Spanish, but the test only measured reading in Spanish, the construct of Spanish proficiency would be underrepresented by the assessment. *Construct-irrelevant variance* refers to a situation in which the test measures other constructs that are irrelevant to the intended construct. Examples of construct-irrelevant variance undermining test score interpretations include when some examinees have trouble interacting with the computerized interface of a testing program (i.e., the construct of "computer proficiency" affecting test performance), or when a student becomes overly anxious when taking a test (i.e., the construct of "test anxiety" affecting test performance). The construct of "speed of response" can also represent construct-irrelevant variance if examinees feel rushed to complete a test that is *not* designed to measure how quickly they can answer questions.

For SWD, a disability may interact with the assessment situation to give rise to construct-irrelevant variance in their test scores. Minimizing such irrelevancies is the reason testing agencies provide accommodations to the standardized testing situation for SWD. Thus, the purpose of test accommodations is to allow students to demonstrate their performance in a manner such that confounding factors related to their disability or language proficiency are minimized. The logic is that accommodations will remove, or at least reduce, any obstacles inherent in a standardized testing situation that will prevent an examinee from demonstrating their proficiency with respect to the construct measured. However, concerns remain that providing an accommodation may change the construct measured, and, in some cases, that change may make the test easier for examinees who receive an accommodation. For example, if the ability to respond to test questions quickly is part of the construct of a reading fluency test, and some examinees get extra time to take the test, the construct may change from reading fluency to reading comprehension, and answering the questions based on only comprehension, rather than fluency, may provide an advantage to those who get extra time. For this reason, the effect of test accommodations on the constructs measured by a test is a critical validity issue and directly affects score comparability across standard and accommodated tests.

### **Accommodations Versus Designated Supports**

For some statewide testing programs, there is also a distinction between "accommodations" and "designated supports." The latter term refers to supports during the testing process that can be given to students with or without disabilities, as determined by local educators, even if the student does not have an IEP or a 504 plan. That is, accommodations are specified in the IEP or 504 plan for a student, whereas designated supports can be prescribed for both general education students and SWD. In this chapter, we use the term "accommodation" to apply to any type of support that represents a change to standard testing conditions.

### Accommodations Versus Modifications

There is an important distinction between accommodations thought to affect the construct measured by a test and those that are thought to leave the construct unaltered. Testing accommodations refer to changes in the test or test administration condition that are *not* thought to alter the construct measured. Thus, test accommodations attempt to remove construct-irrelevant barriers to students' test performance while maintaining construct representation. Changes that *are* thought to alter the construct measured are referred to as *modifications* (AERA et al., 2014). When a test is modified, the scores from original and modified versions of the test are considered to be too different to be comparable; in fact, they are considered to be scores from two different tests. Scores from the standard test given with accommodation are most often considered comparable to those from the test given under standard conditions. The classification of a change to standard testing conditions as an accommodation or a modification is based on both theory and research. Research related to the comparability of scores from standard and accommodated test administrations is discussed later in this chapter.

### Validity and Fairness

A concept closely related to (but broader than) validity is "fairness." This concept is important in test accommodations that are seen as promoting fairness in testing. Moreover, "unfairness" in the provision of accommodations becomes an issue when examinees who take a test with or without accommodations are competing for a common goal (e.g., college admission). Thus, it is important to discuss issues of fairness in testing as they relate to test accommodations.

The AERA et al. (2014) *Standards* state that fairness in testing is "a fundamental validity issue and requires attention to detail throughout all stages of test development and use" (p. 49). The *Standards* also define two related concepts involved in fairness: accessibility and universal test design (UTD). *Accessibility* is defined as "the notion that all test takers should have an unobstructed opportunity to demonstrate their standing on the construct(s) being measured" (p. 49). UTD is defined as "an approach to test design that seeks to maximize accessibility for all intended examinees" (p. 50).

A test that is fair within the meaning of the *Standards* reflects the same construct(s) for all test takers, and scores from it have the same meaning for all individuals in the intended population; a fair test does not advantage or disadvantage some individuals because of characteristics irrelevant to the intended construct. (p. 50)

Another way of saying that scores from a test "have the same meaning for all individuals" is to say scores from the same test are "comparable" across all examinees. In this chapter on fairness, the *Standards* describe "comparability of scores" as enabling "test users to make comparable inferences based on the scores for all test takers" (p. 59).

Mislevy et al. (2013) describe how test accommodations can improve validity in educational testing. As they put it, "making assessment accessible to a more diverse population of students highlights situations in which making tests identical for all examinees can make a testing procedure less fair: Equivalent surface conditions may not provide equivalent evidence about examinees" (p. 121). As they describe, forcing

standardized testing conditions to hold for all examinees will lead to reduction in measurement accuracy for examinees for whom the conditions inhibit their performance.

*Universal test design* refers to a principle of test development whereby the needs of SWD and ELs are considered while the tests are being constructed. The goal of UTD is to make the test and testing situation flexible enough so that accommodations are not necessary (Thompson, Blount, & Thurlow, 2002; Thurlow, Lazarus, Christensen, & Shyyan, 2016). Essentially, UTD calls for test construction practices that focus on eliminating construct-irrelevant variance and more flexible test administration conditions that would make the provision of test accommodations for SWD and ELs unnecessary. For example, removing time limits on a test not only benefits students who have information processing disabilities; it has the potential to benefit all test takers. Elliott and Kettler (2016) provide specific examples of how UTD principles can be used in test development to increase access to educational assessments. Many of these examples reflect sound test development practices, such as reducing unnecessary language load and use only plausible distractors (incorrect response options) on multiple-choice items.

### TYPES OF TEST ACCOMMODATIONS

Test accommodations for SWD can be classified into five categories: (1) timing (e.g., providing extra time or alternative test schedules), (2) response (e.g., allowing students alternative ways to respond to the test, such as using a scribe), (3) setting (changes to test setting), (4) presentation (alternative ways to present test materials), and (5) equipment and devices (allowing students to use additional references or devices). Within each category, there can be many variations. For example, in many statewide educational testing programs, states provide more than 30 separate types of accommodations to students with disabilities who qualify and are approved for the accommodation. In Table 7-1, we provide an example from a mid-sized state. This table provides the frequencies of different types of accommodations on a statewide end-of-course science exam. In this case, 33 separate accommodations were provided by the state, with more than 15,000 students (just under 9 percent) receiving an accommodation. We also break down the groups of students by ELs and non-ELs. For both ELs and non-ELs the most frequent accommodation was a small-group setting. For non-ELs, the next most frequent was extended time, which was the third most frequent accommodation for ELs. The read-aloud accommodations were also relatively frequent for both groups. The percentage of students receiving an accommodation was much higher in the EL group (about 30 percent), due to eligibility for the Spanish version of the exam.

The example presented in Table 7-1 is typical of accommodations provided on most statewide tests that have a sizable population of ELs whose native language is Spanish. Note that accommodations for English learners are covered in Chapter 6 (Comparability When Assessing English Learner Students). Another important point to bear in mind is that many of the students listed in Table 7-1 received more than one accommodation (about 32 percent; see Sireci, Wells, & Hu, 2014). In a later section of this chapter, we review the literature on the effectiveness and validity of presentation, response, and timing accommodations.

**TABLE 7-1** Accommodation Frequencies on a Statewide Science Test

Accommodation Type	Accommodation Condition	Sample Sizes		
		Non-EL	EL	Total
Setting	Small-group testing	10,611	508	11,119
	One-on-one	258	16	274
	Hospital/home testing	62	0	62
	Other setting	93	12	105
Presentation	Some items/questions read aloud	784	100	884
	All items/questions read aloud	785	30	815
	Audio	723	11	734
	Spanish version	15	481	496
	Items interpreted for ELs	3	68	71
	Large print format	59	0	59
	Items signed	12	0	12
	Braille format	7	0	7
	Other presentation	19	5	24
Equipment and Devices	Color chooser	40	0	40
	Magnification device	13	1	14
	Amplification device	10	1	11
	Electronic screen reader	2	0	2
	Color overlay	2	0	2
Timing	Extended time	5,508	309	5,817
	Frequent breaks	1,261	21	1,282
	Changed test schedule	250	14	264
	Other timing	30	1	31
Response	Translation dictionary for ELs	0	117	117
	Administrator transcribed answers	131	2	133
	Administrator marks CR answers	88	0	88
	Administrator marks MC answers	63	1	64
	Typewrite, word processor, computer	42	0	42
	Interpreter for ELs	0	2	2
	Notetaker	2	0	2
	Interpreter translated response	0	1	1
	Augmented communication device	0	0	0
	Other response	22	4	26
Students receiving an accommodation		14,406	1,117	15,523
Students NOT accommodated		162,813	2,675	165,488
Total students		177,219	3,792	181,011
% Accommodated		8.1%	29.5%	8.6%

There are two multistate assessment consortia in the United States that are also illustrative of the types of accommodations provided to schoolchildren on educational tests. The Smarter Balanced Assessment Consortium and the Partnership for Assessing Readiness for College and Careers (PARCC) are multistate assessment consortia that represent groups of states working together to deliver common assessments in reading and mathematics for elementary, middle, and high school students.

PARCC and Smarter Balanced use different terminology for the same accessibility features provided in their assessment systems. For example, Smarter Balanced uses “universal tools” (for accessibility features available to all students), “designated supports” (for accessibility features available upon recommendation from an adult), and accommodations (for accessibility features available to SWD, ELs, and EL SWD based on one’s IEP and 504 plan). PARCC, on the other hand, uses “features for all students,” “accessibility features,” and “accommodations,” respectively. A summary of these accessibility features is presented in Table 7-2.

### **Accommodations on College Admissions Tests**

There are two main college admissions tests in the United States: the SAT and the ACT. Although these tests are similar in many ways, they differ somewhat with respect to their accommodations policies. The College Board, which owns the SAT, provides a list of some of the accommodations they offer on their website (College Board, 2019). The most common accommodations listed are extended time, computer use for essays, extra and extended breaks, reading and seeing accommodations, and use of a four-function calculator. In addition to these, students can request others in their application, which is required to receive accommodations on the SAT. For a student to be eligible to receive accommodations on the SAT, they need to have a documented disability that affects participation in the SAT, and they need to demonstrate that the accommodation(s) they are requesting are necessary. Most students who receive an accommodation on the SAT also receive that accommodation on school tests. However, not all accommodations provided on school tests are allowed on the SAT—they must still be approved by the College Board. Along the same lines, the College Board website does not specify that a student needs to have an IEP or 504 plan, but does specify that there needs to be documentation of the accommodation being received in school, which typically only occurs when a student is on one of those two plans.

The ACT differs in that it does not list possible accommodations for the test on its website; however, it provides a detailed application form for requesting any accommodation. Depending on the accommodation needed, the location of the test may be different. There are two types of testing locations, “national” and “special,” with the most common accommodations being offered at the national locations, but not more specific accommodations. Students may apply for specific accommodations on the ACT; however, to be eligible to receive accommodations students need to provide documentation proving a requirement, such as an IEP, a 504 Plan, or a recommendation from a diagnosing professional (ACT, 2019).

In the 2017–2018 school year, 16 states used the ACT as a required test for graduating students (ACT, 2018), and some other states required the SAT. Requiring students

**TABLE 7-2** Embedded Accessibility Features Provided by Smarter Balanced and PARCC

Accessibility Features and Target Group	Accommodation Category	Smarter Balanced	PARCC
Universal Tools/ Features for All Students	Response	Calculator, digital notepad, highlighter, writing tools	Eliminate answer choices, highlight tool, writing tools
	Presentation	Zoom, strikethrough, spell check, English dictionary, English glossary, expandable passages, global notes, keyboard navigation mark for review, math tools	Audio amplification, bookmark, head phones/noise buffers, line reader mask tool, notepad, pop-up glossary, magnification/enlargement device, spell check
	Time/scheduling	Breaks	
	Presentation	Color contrast, masking, text-to-speech, translated test directions, translations (glossary), translations (stacked), turn off any universal tools	Answer masking, color contrast, text-to-speech
Accommodations for SWD, EL with SWD, EL	Response	Text-to-speech	Text-to speech, grade level calculator
	Presentation	American Sign Language, Braille, closed captioning, streamline	Closed captioning of multimedia (video) for ELA/literacy, American Sign Language video for ELA/literacy, American Sign Language video for mathematics assessments, online transadaptation of the math assessment in Spanish

SOURCES: PARCC (2016); Smarter Balanced (2016).

to take these tests is controversial (see NCME, 2019) and can complicate the accommodations students receive (Lazarus & Thurlow, 2016). For ACT tests taken as a state requirement, the accommodation policies differ from the policies for taking the ACT as a college admissions test. Thus, there are three types of accommodations possible: (1) college-reportable accommodations, (2) state-allowed accommodations, and (3) local arrangements. College-reportable accommodations are approved by ACT. State-allowed accommodations are not approved by ACT, but the state allows them for a particular student. These scores are reported only to the school for state assessment use, are not stored in the ACT database, and are not able to be sent to colleges or scholarship agencies. Local arrangements use private room or small-group accommodations and ACT approval is not needed, thus allowing the scores to be college reportable (Lazarus & Thurlow, 2016).



The accommodation policy of the SAT differs from that of the ACT for states using the assessment as a requirement. Lazarus and Thurlow (2016) compared SAT policies used in three different states: Connecticut, Michigan, and New Hampshire. New Hampshire and Michigan provided example accommodations, but there was no set list of accommodations available to students. The states required students to submit an IEP or 504 plan showing a functional impact to be qualified for accommodations. The College Board ultimately decided which accommodations were allowed for each student who applied. In Connecticut, the state and the College Board jointly decided which accommodations were appropriate for students. They included a list of accommodations that were allowed for the scores to be college reportable, and a list of those that were not, but the decision ultimately came down an agreement between the state and the College Board (Lazarus & Thurlow, 2016).

Prior to 2003, when students took the ACT or the SAT with an accommodation, their score report was “flagged” with a footnote stating that the test was taken under nonstandard conditions. Although these footnotes are intended to point out the potential noncomparability of scores from accommodated test administrations and standard administrations, disability rights groups successfully argued that such warnings essentially informed college admissions officers that a student had a disability and were therefore discriminatory. In 2003, both the College Board and the ACT ended the practice of flagging scores on accommodated tests. Since that time, articles have appeared in the popular press suggesting that more affluent parents are more frequently receiving disability diagnoses for their children so they can get extra time accommodations on these tests. For example, *The New York Times* reported an increase of accommodated administrations of the SAT from 2 percent in 2002 to 4 percent in “recent years” (Goldstein & Patel 2019). There are similar concerns raised for accommodations granted on other admissions tests, such as those for high school. Clearly, false diagnoses of a disability for the purpose of gaining an advantage on a test is criminal behavior. However, such increases in accommodations could be occurring for other reasons: (a) improved diagnoses of disabilities, which more affluent families are more likely to have access to, and (b) less stigmatization of SWD who apply for accommodations. Either possibility suggests an improvement in the validity of these scores (i.e., a valid accommodation for a valid disability). Nevertheless, future research should be done to shed light on this issue, and efforts should be made to decrease the “diagnosis gap” between higher and lower socioeconomic groups. Recent research has also indicated disparities in special education diagnoses by race (Elder, Figlio, Imberman, & Persico, 2019; Shifrer & Fish, 2019; Zerkel & Weathers, 2016), which suggests additional work is needed to ensure valid assessment and diagnosis of potential disabilities for all students.

### COMPARABILITY ISSUES IN TEST ACCOMMODATIONS

As mentioned earlier, a goal of standardized testing is to provide scores that can be interpreted in the same way for all individuals who take the same test. Many physical measurements, such as weight and height, can be compared across people, as long as we are using the same scale (e.g., pounds or kilograms, feet or meters). However, test score scales in educational and psychological testing, even when standardized, may

have issues that threaten score comparability across examinees. Clearly, if SWD are forced to take a test under conditions in which their disability impedes their performance, their scores cannot be interpreted in the same way as an examinee who does not have the same impediment. Similarly, changing the standardized conditions of an assessment can lead to a lack of comparability across scores from accommodated and standard test administrations.

In situations involving test accommodations it is generally accepted that scores from standard and accommodated tests cannot be “equated” in the traditional sense. Therefore, research is often conducted to evaluate whether accommodations may alter the construct measured, change the interpretation of a test score, or provide an advantage or disadvantage to SWD or examinees without disabilities.

### Evaluating Comparability

Earlier, we mentioned that the AERA et al. (2014) *Standards* specify five sources of validity evidence that can be used to evaluate the use of a test for a particular purpose. Questions regarding the comparability of scores across standard and accommodated tests are essentially validity questions: Can scores from standard and accommodated test administrations be interpreted and used in the same way?

Although all five sources of validity evidence are germane to evaluating test accommodations, with respect to evaluating the comparability of scores from accommodated and standardized tests, two sources are particularly relevant: validity evidence based on relations to other variables and validity evidence based on internal structure. Accommodation studies that fall into the category of validity evidence based on relations to other variables are studies that have evaluated the interaction hypothesis (NRC, 2004) and the differential boost hypothesis (Elliott & Kettler, 2016; Fuchs, Fuchs, Eaton, Hamlett, & Karns, 2000). As we explain, these studies treat groups of students defined by accommodation and disability status as the external variables under study.

The *interaction hypothesis* states that, when test accommodations are given to SWD who need them, their test scores will improve relative to the scores they would attain from taking the test under standard conditions, and students without disabilities will *not* exhibit higher scores when taking the test with an accommodation. Thus, the interaction specified is between student group (SWD or non-SWD) and test administration condition (accommodated versus standard). This interaction can be considered in the context of a factorial design where a within-subjects factor (standard or accommodated test administration) interacts with a between-subjects factor (student group).

The *differential boost hypothesis* is similar but represents a more realistic depiction of the effectiveness of test accommodations by relaxing the hypothesis that students without disabilities will not have score gains in the accommodation condition. According to the differential boost hypothesis, if an accommodation is effective, the gains for SWD will be greater than the gains observed for non-SWD. Like the interaction hypothesis, differential boost is evaluated using experimental designs where one factor is the student group and the other factor is the test administration condition. These studies typically use analysis of variance to evaluate the main and interaction effects. Reviews of differential boost and interaction hypothesis studies for SWD can be found

in Kettler (2012) and Sireci, Scarpatti, and Li (2005). The findings generally support the differential boost hypothesis in that SWD generally benefit more from accommodations relative to students without disabilities. In the next section of this chapter, we review research in this area.

Validity evidence based on internal structure has been used to evaluate the comparability of accommodated and standard test scores by evaluating whether the dimensionality, or “structure,” of an assessment is consistent across students who received an accommodation and students who did not. Analysis of *differential item functioning* (DIF), which investigates whether the statistical characteristics of an item are inconsistent across groups (in this case across standard and accommodated test conditions), also falls under the internal structure category. Other examples of statistical procedures used to evaluate the equivalence of a test across groups of students defined by SWD or EL status include item response theory, structural equation modeling, and multidimensional scaling.

Regardless of the source of validity evidence used to evaluate comparability, it is important for the comparability analysis to focus on the purpose of the test. For example, if the purpose of a test is to determine whether a student has mastered the material taught in a sixth grade math class, the comparability of the *scores* from accommodated and standard test administrations is relatively unimportant, whereas the comparability of the “master” versus “did not master” *decision* based on those scores is of the utmost importance. In this example, the test is used to determine mastery, and so that decision across the standard and accommodated test version needs to be validated. However, if a test is used to rank-order examinees, as when students are competing for a limited number of scholarships, the test scores themselves are used in making the rankings, and so the comparability study should focus on the test score level. In the next section, we review some comparability studies that have been conducted to illustrate recent research on evaluating test accommodations for specific purposes.

### REVIEWING THE LITERATURE ON EVALUATING TEST ACCOMMODATIONS

Many studies have been done to evaluate the effects of test accommodations on the validity, interpretability, and comparability of test scores. In this section, we review some of this literature, focusing on meta-analyses that analyzed the results from multiple studies. Our review is organized by the five most common types of accommodations that have been studied: timing, presentation (which is primarily the read-aloud accommodation), response, setting, and equipment and devices. However, as mentioned earlier, students who receive accommodations often receive more than one type of accommodation, and some of the studies reviewed looked at combinations of accommodations. The research in this area has looked at both validity evidence based on relations to other variables (specifically, the interaction and differential boost hypotheses), and validity evidence based on internal structure (i.e., consistency of factor structure and item functioning across standard and accommodated test administrations).

### Timing Accommodations

Timing accommodations include alterations to the time limit or time spent taking a test such as provision of extra time and alternative test schedules. Gregg and Nelson (2012) conducted a meta-analysis on studies investigating the provision of extra time to students with learning disabilities. They reviewed nine studies that met the criteria for their meta-analysis, which involved comparisons across students with learning disabilities and students without learning disabilities who took reading or math exams with or without extra time. The studies were published between 1986 and 2007. Although they found some evidence for the differential boost hypothesis in that students with the most severe learning disabilities had higher gains with extra time than students without learning disabilities, they concluded that the results across studies led to more questions than answers. As they put it, "The literature is lacking in quantity of studies, restricted in types of design methodologies, and under representative of the diversity of individuals demonstrating the disorder" (p. 134).

Lovett (2010) also reviewed the literature on extra time accommodations. He looked at 20 empirical studies to investigate whether extra time would change the construct measured or would benefit students without disabilities. With respect to studies that used factor analysis to evaluate whether the factor structure was the same across tests given with and without extra time, the results generally support comparability. However, for studies that looked at the predictive validity of the test scores (e.g., if scores from tests given with extra time correlate with college grades to the same extent as scores from tests given without extra time), the results indicated that law school admissions test scores and SAT test scores from tests administered with extended time had less predictive validity than tests administered with standard time. However, Lovett pointed out the nonexperimental nature of these studies (see also Sireci [2005], regarding problems with predictive validity studies for SWD).

With respect to whether students without disabilities would benefit from extended time, Lovett found that, similar to Sireci et al. (2005), students both with and without disabilities tended to benefit from extended time, with some evidence that SWD had greater benefits. He concluded that "there is at least some evidence supporting the differential boost hypothesis, for tests that are not highly speeded" (p. 624).

Turning from the more general reviews on extended time to some of the specific studies, it is worth noting that Cohen, Gregg, and Deng (2005) evaluated extra time accommodations on a statewide test for ninth graders at both the item and test score levels. Using differential item functioning analyses, they concluded that items functioned similarly across groups defined by accommodation and disability status. In addition, they found that SWD and general education students both showed improvement in performance with extra time, indicating that both groups benefited from the accommodation without differential boost. The authors noted that this is an indication that perhaps the accommodation should be considered for all students.

Fletcher et al. (2006) compared the effects of extra time on the test scores of students with and without reading disabilities. The students with reading disabilities all exhibited poor word-decoding skills, while the students without reading disabilities all scored in the average range for word-decoding skills. In their study, students received three accommodations (double time, proper nouns read aloud, and reading comprehension stems read aloud), or no accommodations. It is more difficult to isolate the effect of

extra time from the effects of the other two accommodations; however, they found that students with poor word-decoding skills showed differential boost over the students with average word-decoding skills (the effect size was about a 0.75 standard deviation). These findings indicate the accommodations were useful to all students, but especially to those who had difficulty accessing the assessment.

In another nonexperimental study, Searcy, Dowd, Hughes, Baldwin, and Pigg (2015) found that students who were granted an extra-time accommodation on the Medical College Admissions Test had essentially equal rates of acceptance into medical school as students who did not have extra time. However, they also found that the students who received extra time had lower rates of passing the U.S. medical licensure exam within 8 years of completing medical school.

To summarize the results of extra-time accommodations, the literature indicates positive effects for all students, which as Sireci et al. (2005) noted may be due to unintended speededness in tests. However, the evidence also suggests SWD benefit more from extra time than students without disabilities, which adds support for providing this accommodation when it is needed.

### **Read Aloud as a Presentation Accommodation**

The most common accommodation with respect to how test material is presented is the read-aloud accommodation. The implementation of this accommodation can vary between computer platforms, tests, students, and jurisdictions; however, it typically means that some part of the test (items, passages, or both) is being orally presented to the student via another person (such as a test proctor) or a computer (e.g., screen-reading software).

Buzick and Stone (2014) conducted a comprehensive meta-analysis on the effects of read-aloud test accommodations across 19 studies conducted between 1998 and 2013. Of the 19 studies, 11 focused on math tests, 6 focused on reading tests, and 2 focused on both. As in most other reviews on test accommodations, they did not find support for the interaction hypothesis, but found support for the differential boost hypothesis on reading tests in that SWD generally exhibited differential boost under the read-aloud condition relative to students without disabilities. The boost was greater in elementary schools than in high schools but was not prominent for math tests.

Li (2014) used meta-analysis to analyze 114 effect sizes reported in 23 studies that investigated the effects of read-aloud accommodations on reading and math tests. She used hierarchical linear modeling to look at the overall effects, as well as effects moderated by grade level, subject area, delivery method (human reader, audiocassette, or computer), and research design (independent groups versus repeated measures). Similar to the analysis by Buzick and Stone (2014), she found an overall differential boost effect for SWD of about a 0.20 standard deviation, with the effect being larger for reading tests than math tests. Larger effects were also found for read-aloud accommodations delivered via a human reader, and for elementary school grades. She noted that, because there were overall gains for both SWD and students without disabilities, more research is needed to determine whether the read-aloud accommodation should be restricted to only SWD.

Buzick and Stone's (2014) and Li's (2014) meta-analyses both found that the read-aloud effect is larger for reading than for mathematics for all students, and that the



effect is larger for SWD than for students without disabilities. Additionally, both studies found the effect sizes were larger for elementary school grades than for middle school grades. The two studies were also similar in that they recommended further research to determine whether it is fair to provide read-aloud accommodations to only SWD.

In addition to focusing on the differential boost effects of read-aloud accommodations, some studies have investigated whether this accommodation changes the construct measured. For example, Huynh and Barton (2006) looked at whether a read-aloud accommodation affected the factor structure of a statewide high school reading test. They compared the factor structure of the test across three groups: SWD who took the test with a read-aloud accommodation, SWD who took the test without a read-aloud accommodation, and students without disabilities who took the test without a reading accommodation. They found that the factor structure was the same across all three groups, which indicated the read-aloud accommodation did not change the reading construct measured. They also found that SWD who took the test with the read-aloud accommodation performed similarly to SWD who did not receive the accommodation. They assumed the latter group had less severe disabilities and so concluded that “oral administration accommodation served to level the playing field for students whose disabilities were presumably severe enough to require oral accommodations” (p. 21).

Cook, Eignor, Steinberg, Sawaki, and Cline (2014) also used factor analysis to evaluate whether a read-aloud accommodation on a reading comprehension test changed the construct measured. Using both confirmatory and exploratory factor analysis methods, they compared the factor structure across four groups: SWDs with reading disabilities who took the test with or without the read-aloud accommodation, and students without disabilities who took the test with or without the accommodation. The accommodation condition was randomly assigned to students. Their results supported invariance of the factor structure across all four groups, which led them to conclude the test was measuring the same construct across the read-aloud accommodation and the standard administration.

### **Response Accommodations**

A response accommodation refers to nonstandard ways that students communicate their answers. Bolt and Thurlow (2004) completed a synthesis of the research on the five most frequently allowed accommodations in state policies, one of which was dictated response to a scribe. This accommodation was allowed in 37 states at the time of publication. Three of the studies used differential boost to determine the effect of the accommodation on SWD versus students without disabilities. All three found that SWD had greater gains under the dictation condition than students without disabilities. However, the differential boost was statistically significant for only two of the three studies. They also noted that two studies looked at length and quality of writing passages, one of which found the writing was significantly longer for SWD who dictated their response; the other study found no difference. However, only one study in their review compared computerized speech to text versus teacher scribing. With the increase in computerized testing, it is likely that this accommodation will be more technology based in the future.

Another type of response accommodation is to allow students to write their answers directly in a test booklet rather than filling in the answer sheet. Fuchs, Fuchs, and



Capizzi (2005) mention that there is no empirical research to support this accommodation, although it is used frequently and is inexpensive. They mention a few studies that have shown no statistically significant results for either SWD or students without disabilities.

### **Setting Accommodations**

A setting accommodation is one where a student takes the assessment in a separate room or building, often one on one with a proctor or in a small group. These accommodations are typically thought to address attention problems that some students with learning disabilities have by placing students in a less distracting environment. Setting accommodations are frequently combined with other accommodations, such as read-aloud accommodations or extra time, so the student's accommodations will not interrupt other students taking the assessment in the general classroom.

Lin and Lin (2014) investigated whether a setting accommodation could account for differential item functioning on a test involving sixth grade students with learning disabilities in both math and reading. Students with learning disabilities were compared to students without. Other factors such as language spoken at home and confidence in the areas being tested were also investigated as covariates. Three items on the math assessment were flagged for "moderate" DIF, and six items were flagged for moderate DIF on the reading assessment. For both subject areas, students scored better on only half the items (one math and three reading) under the setting accommodation condition. To explore these findings, they analyzed students' background variables and concluded it is not possible to know whether it was the setting accommodation or student background characteristics that explained these differences.

Lewandowski, Wood, and Lambert (2015) also investigated the effect of a setting accommodation—specifically, a private room. However, SWD were not identified as a separate group. Instead, the focus was on whether there was any effect for a "typical" group of college students. Two parallel forms of a reading comprehension test were administered to 62 students, with the setting accommodation condition counterbalanced. Each student took one form in a group setting and one form in a private room. They found students' scores were higher in the group setting condition, and on the second testing (suggesting a practice effect). They concluded that because a private room did not provide an advantage for the students, it should be considered a valid test accommodation. However, Lewandowski et al. also recommend a follow-up study that includes SWD to see if they benefit from the accommodation.

It appears very few empirical studies on setting accommodations have been conducted. In their synthesis of the most frequently allowed testing accommodations in state policies, Bolt and Thurlow (2004) hypothesized this lack of research may be due to the fact that setting accommodations are not seen as controversial, given there is no reason to suspect a change in the construct measured, or a legitimate advantage.

### **Equipment and Devices Accommodations**

An equipment and materials accommodation typically refers to the use of additional references or devices. Engelhard, Fincher, and Domaleski (2011) used a repeated-measures design to evaluate the effects of two accommodations: calculators and resource

guides. They were curious whether these accommodations affected students with or without disabilities, and if the two groups were affected differently. The sample used was deliberately chosen to match Georgia school demographics, including students with each specific disability type. They found that resource guides were not helpful for students with or without disabilities. In fact, student performance declined with the use of resource guides. However, they noted resource guides could be configured differently by states. They also found that calculators helped student performance, but the increase was larger for students *without* disabilities than for those with disabilities. That finding was similar to that of Fuchs, Fuchs, Eaton, Hamlett, Binkley, and Karns (2000), who found that the performance of students without learning disabilities improved more than students with learning disabilities with the use of a calculator. Based on these studies, it appears that the use of calculators does not achieve the goal of helping the students who need them better access the test.

The use of calculators on standardized testing has evolved over the past 10 years, but research on the accommodation use has not yet been updated. Further research should be done in this area, especially focusing on calculator use in computer-based assessments.

### Summary of the Review

Although our review of the literature was not extensive and focused primarily on previous reviews, the conclusions that can be drawn are that for both extended-time and read-aloud accommodations, there is evidence of differential boost for SWD. These accommodations tend to improve the performance of SWD more than that of students without disabilities. However, this finding is not consistent across all studies, and the fact that students without disabilities experience some improvement in performance suggests that standard testing conditions should allow these accommodations for all students, particularly when speed of response and reading fluency are not the targets of the assessment. There is also evidence that the factor structure does not change when these accommodations are provided.

With respect to response accommodations, the few studies that have been done suggest that dictating responses and writing answers in test booklets are appropriate accommodations that are not associated with increases in students' test performance. This area of research is probably most dated, because answer sheets and scribes may become obsolete with the technology tools offered via computer-based testing. The research suggests calculators are beneficial to both SWD and students without disabilities, but more research is also needed in this area.

The trends we identified in our review suggest some directions for future research and practice. However, a much more extensive set of research-based guidelines for test accommodations was provided by Abedi and Ewers (2013), which we turn to next.

### Deriving Recommendations from the Literature

In a commissioned study by the Smarter Balanced Assessment Consortium, Abedi and Ewers (2013) convened a panel of five experts on accommodations for both SWD and ELs to conduct a systematic review of accommodations research and provide

recommendations regarding the appropriateness and validity of different types of accommodations.

Based on a review of the literature and their expertise, each panelist rated each accommodation on two dimensions: (1) whether the accommodation would alter the construct measured by the test, and (2) whether the accommodation would make the test more accessible for the students who would need it. They also determined whether each accommodation might improve the performance of *all* students (not just ELs) in a way that would not affect the construct. If so, they listed the accommodation as “access,” and concluded the accommodation improved access for all students by reducing construct-irrelevant variance. In Table 7-3 we present a summary of Abedi and Ewers’s final ratings for the accommodations for SWD.

We do not consider the recommendations presented in Table 7-3 to be absolute or authoritative. However, they provide a good summary of the literature up to 2012, with respect to specific types of accommodations and their appropriateness for specific testing situations (see also Elliott & Kettler [2016] for guidance in this area). Thus, these recommendations provide a good starting point for testing agencies and test users to determine which specific accommodations may be appropriate for specific types of students in specific testing situations.

### ALTERNATE ASSESSMENTS FOR STUDENTS WITH SEVERE DISABILITIES

As mentioned earlier, states have created *alternate assessment* systems for SWD with significant cognitive disabilities who require more than just accommodations to standard testing conditions to allow them to fully participate in educational assessments. Students who require alternate assessments have intellectual disabilities such as severe forms of autism or multiple cognitive impairments. Currently, the U.S. Department of Education allows a maximum of only 1 percent of public school students in a state to take alternate assessments used for accountability purposes.

Alternate assessments are designed for students who are unable to take the regular assessment, even when given accommodations. Typically, these assessments differ in content from the general assessments in that they measure “extended” or “alternate” content standards that are in some way aligned with the overall state standards (DOEd, 2015; Kearns, Towels-Reeves, Kleinert, Kleinert, & Kleine-Kracht, 2011). Students with significant cognitive disabilities often need adaptations, scaffolds (i.e., assistance by test administrators or the computer, which are removed slowly as the student’s competency increases; see Azevedo & Hadwin, 2005), and supports to access age- and grade-appropriate curriculum content.

Students with significant cognitive disabilities often use augmentative and alternative communication devices in school settings because they have difficulty in expressive and receptive communication. Such devices include all forms of communication (other than oral speech) that are used to express thoughts, needs, wants, and ideas. The accessibility assessment features suggested for these students are primarily technology based and include answer masking, audio players, line readers, magnification, inverted color choice, color contrast, overlay color, read aloud with highlighting, text to speech, uncontracted Braille, sign interpretation of text, language translation of text, scribing,

**TABLE 7-3** Summary of Recommended Accommodations for SWDs from Abedi and Ewers (2013)

Accommodation	Risk
Test administration directions simplified or clarified (does not apply to test questions)	None
Large-print versions/test items enlarged if font larger than required on large-print versions	None
Customized dictionary/glossary (content-related terms removed)	None
Pop-up glossary Computer Based Testing (CBT) (content-related terms excluded)	None
Computer use (including word processing software with spell- and grammar-check tools turned off for essay responses to writing portion of a test)	None
Calculator on mathematics tests (if not part of the focal construct)	None
Calculator on science tests (if not part of the focal construct)	Minor
Test questions read aloud to student	Minor
Manually coded English or American Sign Language to present directions for administration	Minor
Manually coded English or American Sign Language to present test questions	Minor
Braille transcriptions provided by the test contractor	Minor
Audio amplification equipment	Minor
Colored overlay, mask, or other means to maintain visual attention	Minor
Special lighting or acoustics; special or adaptive furniture such as keyboards, larger/antiglare screens	Minor
Visual magnifying equipment	Minor
Assistive device that does not interfere with the independent work of the student	Minor
Arithmetic table or formulas (not provided) on the <i>mathematics</i> tests if not part of the focal construct	Minor
Math manipulatives on <i>science</i> tests (if they do not interact with intended construct)	Moderate
Math manipulatives on <i>mathematics</i> tests (if they do not interact with intended construct)	Moderate
Arithmetic table or formulas (not provided) on <i>science</i> tests if not part of the focal construct	High

NOTES: "Risk" describes the extent to which the accommodation was judged to possibly change the construct measured. "None" means Abedi & Ewers (2013) did not list a level of risk associated with the accommodation.

SOURCE: Adapted from Abedi & Ewers (2013).

and many others (see Lee, Browder, Wakeman, Quenemoen, & Thurlow, 2015; Wells-Moreaux, Bechard, & Karvonen, 2015).

Two prominent alternate assessment systems in the United States are Dynamic Learning Maps (DLM) and the National Center and State Collaborative, the latter of which is now referred to as the Multi-State Alternate Assessment (MSSA). These two systems represent groups of states (consortia) that have come together to develop common assessments for students with severe cognitive disabilities.

Both DLM and MSSA begin by transforming general curriculum standards, such as the Common Core College and Career Readiness Standards, to alternate assessment standards using appropriate adaptations, scaffolds, and supports. These alternate assessment standards represent the general intent of a curriculum standard in a way that is more appropriate for these students' cognitive functioning and instructional experiences. For example, DLM's "essential elements" are derived from the college and career readiness standards and are aligned to grade level, but at reduced depth, breadth, and complexity (Wells-Moreaux et al., 2015). MSSA's "alternate achievement standards" are based on an adapted general age- and grade-appropriate academic content (Herrera, Turner, Quenemoen, & Thurlow, 2015).

Using the Universal Design for Learning framework, DLM claims to make its assessments more accessible by providing communication and alternate access tools for students to use in the classroom (e.g., communication boards, alternate keyboards). DLM uses an adaptive testing system in that it administers an initial set of test items (module) to all students at the beginning of the assessment to determine students' ability levels. Using this information, the DLM assessment system routes the students to an appropriately challenging subsequent sets of tasks that closely match their knowledge and skills in grade-level essential elements (Clark, Kingston, Templin, & Pardos, 2014). Students are given sets of reading and writing items called "testlets." In mathematics, these testlets are either multiple-choice or technology-enhanced items (e.g., the computer interface allows students to use graphics to display or provide new information when providing responses). The tests may be taken independently using accessibility features like alternate keyboards, touch screens, or switches or with support from a test administrator depending on each student's information from the IEP, the educator, and the first contact survey information (Wells-Moreaux et al., 2015).

The MSSA system is designed so that students begin with less complex test items and with more adaptations, scaffolds, and supports; then students move to more complex test items with reduced supports, as appropriate. In elementary mathematics, the assessment concentrates on number operations relations, spatial relations, and measurement; for middle and high school, the system concentrates on problem solving with supports that may include definitions and demonstrations. For English language arts (ELA) the system assesses reading (which may be verbal or nonverbal), comprehension, and writing, using supports and scaffolds like introduction to text, rereading, pictures, definitions, and prompts for what to listen for (Herrera et al., 2015). The system mainly uses computer-administered selected-response items for both ELA and mathematics and a few constructed-response items in both subjects. Most adaptations and supports are built into the system and are computer delivered (e.g., alternate keyboard, switches, and hubs), but some may be provided by humans (e.g., a scribe, sign language test administrator, etc.) (NCSC, 2015).

DLM and MSSA exemplify universal test design for a specific population; that is, accommodations and flexible administrations are part of the standardized assessment protocol. Although these assessment systems are atypical in terms of student population, they may lead the way for more accessible general assessments in the future.

### **Research on Alternate Assessments**

There is relatively little empirical research on alternate assessments, due primarily to the small population size, but some studies have been done. Laitusis, Maneckshana, Monfils, and Ahlgrim-Delzell (2014) investigated DIF on alternate assessments to determine whether certain item characteristics were associated with DIF for specific disability categories. The students in this study were diagnosed with severe cognitive impairments, autism, and orthopedic impairments. Subject-matter experts (SMEs) coded item type based on the state standard being addressed by the item and the types of skills predicted by the SMEs as being necessary to answer the item successfully. They found some item characteristics contributed to DIF. Some of these characteristics, such as rote learning, were construct relevant, but others may represent construct-irrelevant variance (e.g., items that included a social exchange). They suggested evaluating such items to determine whether they can be changed so as not to include irrelevant features that may hinder one group's ability to answer the items correctly.

In another study on alternate assessments, Zebehazy, Zigmond, and Zimmerman (2012) investigated DIF across students with cognitive disabilities and students with both cognitive disabilities and visual impairments. On the test, students without visual impairments were asked to point to the correct picture of an item to answer each question, but students with visual impairments were asked to choose the item by touch (e.g., when asked for something that you use to eat, students without visual impairments pointed to a picture of a fork, while students with visual impairments feel four objects and choose the fork). It was found that the manner of presentation and the reorientation of items is important for students with visual impairment. Additionally, the use of model-sized figures (such as a toy car rather than a picture of a real car) contributed to DIF. It was also found that some of the items that the students with visual impairment found to be more difficult may have resulted from a tendency of the assessment to underaccommodate for that group. Thus, this study has important implications for future alternate assessment development.

### **TEST ACCOMMODATIONS AND TECHNOLOGY**

The computerization of tests provides opportunities to embed supports and other accommodations directly into the testing process in a seamless manner. Simply by making adjustments to a digital device, font sizes can be increased, translated versions of test items can be presented, test content can be provided orally, and voice recognition software can be used to record responses. We are just at the beginning of the era of computer-based test accommodations, but already testing programs are taking advantage of technology to make tests more accessible.

The aforementioned PARCC and Smarter Balanced assessment programs are examples of tests that use technology to improve accessibility for all students. As part of



their federal grants, both testing consortia were required to use principles of universal design for learning to create test items. As discussed previously, they used technology to include many accessibility features and accommodations for students with learning disabilities (see Table 7-2). These technological innovations help foster a more inclusive testing environment, allowing more students to take tests with accommodations in the general education classroom (Batel & Sargrad, 2016).

Audio presentation of test material through technology is a popular example of current technological innovations that improve accessibility. Johnstone, Higgins, and Fedorchak (2019) pointed out that computerized testing allows presentation of read-aloud material in different ways, such as allowing the students to control which parts of a test are read aloud. They evaluated the different ways scientific and mathematical test material was read aloud and concluded that “a comprehensive strategy may be needed to develop scripting rules for assessments” (p. 815), and they argued that audio presentation should be made available to make tests more accessible to all students.

As technology continues to better enable supports and accommodations for educational tests, it will become increasingly important to ensure students know how to use the technology. As with any accommodations, those technological supports students use in the classroom should, as much as possible, match those used on the assessments they take. When they differ, practice tests and clear instructions should be provided to help students take advantage of these tools (Crotts-Roohr & Sireci, 2017).

## CONCLUSION

In this chapter, we reviewed issues, policies, and research related to assessing SWD and the use of test accommodations to provide more valid assessments for these students. Many of the studies we reviewed provided suggestions for future research, such as determining the specific types of accommodations that are best suited to students with particular disabilities, and identifying ways computers can more efficiently and seamlessly make accommodations available to the students who need them.

With respect to improving the congruence between test accommodations and the needs of specific students, more research needs to be done on the *process* of accommodation assignments, and on the knowledge and training of IEP team members and other educators with respect to test accommodations. Test accommodations are often determined based on a student’s IEP or 504 plan, but teachers and other IEP team members may not be familiar with all accommodations offered by a testing program or how they are best implemented. Some research (e.g., Helwig & Tindal, 2003) has shown that teachers’ recommendations for accommodations are not always accurate. For example, they found that about half of the students who were provided accommodations were not helped by the accommodations, and about half of the students who did not receive accommodations would have been helped if they had them. Thus, we recommend including more training for teachers on choosing appropriate accommodations for students within a specific testing situation. Expanding on the guidance provided by Abedi and Ewers (2013) may be helpful in such training. As we better understand how IEP teams and other educators are making accommodation decisions, we can better develop training materials and guidance to maximize the fit between what accommodations are needed for each student, and what accommodations are not.

Training is also likely needed for students. Accommodations are more commonly administered via computers, and there may be gaps in understanding of different types of students in how to navigate a computer interface. Such gaps will interfere with proper use of accommodations. Given that such gaps are likely to be associated with socioeconomic status, immigration status, and other demographic variables, we encourage testing programs to develop, and require, practice tests that include interaction with the accommodation supports. As mentioned earlier, we also need to reduce the “diagnosis gap” that likely exists across different cultural and socioeconomic groups (Elder et al., 2019; Shifrer & Fish, 2019; Zerkel & Weathers, 2016).

Finally, it is important to note that although this chapter discusses SWD, categorizing students into disability categories is an imperfect science, and some (if not all) students *not* classified as having a disability are likely to have deficits that interact with the testing process (e.g., anxiety). Thus, we believe accommodations that do not alter the construct measured should be made available to all students. Further research can help evaluate, and most likely demonstrate, how such universal features will improve assessment validity for all students.

## REFERENCES

- Abedi, J., & Ewers, N. (2013). *Smarter Balanced Assessment Consortium: Accommodations for English language learners and students with disabilities: A research-based decision algorithm*. Smarter Balanced.
- ACT, Inc. (2018). *The condition of college and career readiness*. Retrieved from <https://www.act.org/content/dam/act/secured/documents/cccr2018/National-CCCR-2018.pdf>.
- ACT, Inc. (2019). *Accommodations and English learner supports for educators*. Retrieved from <https://www.act.org/content/act/en/products-and-services/the-act/registration/accommodations.html>.
- AERA (American Educational Research Association), APA (American Psychological Association), & NCME (National Council on Measurement in Education). (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.
- Azevedo, R., & Hadwin, A. (2005). Scaffolding self-regulated learning and metacognition. Implications for the design of computer based scaffolds. *Instructional Science*, 33, 367–379.
- Batel, S., & Sargrad, S. (2016). *Better tests, fewer barriers: Advances in accessibility through PARCC and Smarter Balanced*. Washington, DC: Center for American Progress.
- Bolt, S. E., & Thurlow, M. L. (2004). Five of the most frequently allowed testing accommodations in state policy: Synthesis of research. *Remedial and Special Education*, 25(3), 141–152.
- Buzick, H., & Stone, E. (2014). A meta-analysis of research on the read aloud accommodation. *Educational Measurement: Issues and Practice*, 33(3), 17–30.
- Clark, A., Kingston, N., Templin, J., & Pardos, Z. (2014). *Summary of results from the fall 2013 pilot administration of the Dynamic Learning Maps™ Alternate Assessment System* (Technical Report No. 14-01). Lawrence, KS: University of Kansas Center for Educational Testing and Evaluation.
- Cohen, A. S., Gregg, N., & Deng, M. (2005). The role of extended time and item content on a high stakes mathematics test. *Learning Disabilities Research & Practice*, 20(4), 225–233.
- College Board. (2019). *Accommodations on College Board Exams*. Retrieved from <https://accommodations.collegeboard.org>.
- Cook, L., Eignor, D., Steinberg, J., Sawaki, Y., & Cline, F. (2014). Using factor analysis to investigate the impact of accommodations on the scores of students with disabilities on a reading comprehension assessment. *Journal of Applied Testing Technology*, 10(2), 1–33.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302. <http://doi.org/10.1037/h0040957>.
- Crotts-Roohr, K., & Sireci, S. G. (2017). Evaluating computer-based test accommodations for English learners. *Educational Assessment*, 22(1), 35–53. <http://doi.org/10.1080/10627197.2016.1271704>.

- DOEd (U.S. Department of Education). (2015). *US Department of Education FY2015 annual performance report and FY2017 annual performance plan*. Retrieved from <http://www.ed.gov/about/reports/annual/index.html>.
- DOEd, Office of English Language Acquisition. (2016). *Fast facts: Languages spoken by English learners (ELs)*. Washington, DC: Consolidated State Performance Reports.
- Elder, T. E., Figlio, D. N., Imberman, S. A., & Persico, C. I. (2019, May). *School segregation and racial gaps in special education identification* (NBER Working Paper No. 25829). Cambridge, MA: National Bureau of Economic Research.
- Elliott, S. N., & Kettler, R. J. (2016). Item and test design considerations for students with special needs. In S. Lane, T. Haladyna, & M. Raymond (Eds.), *Handbook of test development* (pp. 374–391). Washington, DC: National Council on Measurement in Education.
- Engelhard, G., Fincher, M., & Domaleski, C. S. (2011). Mathematics performance of students with and without disabilities under accommodated conditions using resource guides and calculators on high stakes tests. *Applied Measurement in Education*, 37, 281–306.
- Fletcher, J. M., Francis, D. J., Boudousquie, A., Copeland, K., Young, V., Kalinowski, S., & Vaughn, S. (2006). Effects of accommodations on high-stakes testing for students with reading disabilities. *Exceptional Children*, 72(2), 136–150.
- Fremer, J., & Wall, J. (2004). Why use tests and assessments? In J. Wall & G. Walz (Eds.), *Measuring up: Assessment issues for teachers, counselors, and administrators* (pp. 3–19). Greensboro, NC: CAPS Press.
- Fuchs, L., Fuchs, D., & Capizzi, A. (2005). Identifying appropriate test accommodations for students with learning disabilities. *Focus on Exceptional Children*, 37(6), 1–8.
- Fuchs, L. S., Fuchs, D., Eaton, S. B., Hamlett, C., & Karns, K. (2000). Supplementing teacher judgments about test accommodations with objective data sources. *School Psychology Review*, 29 (1), 65–85.
- Goldstein, D., & Patel, J. K. (2019, July 30). Need extra time on tests? It helps to have cash. *The New York Times*. Retrieved from <https://www.nytimes.com/2019/07/30/us/extra-time-504-sat-act.html>.
- Gregg, N., & Nelson, J. M. (2012). Meta-analysis on the effectiveness of extra time as a test accommodation for transitioning adolescents with learning disabilities: More questions than answers. *Journal of Learning Disabilities*, 45, 128–138.
- Helwig, R., & Tindal, G. (2003). An experimental analysis of accommodation decisions on large-scale mathematics tests. *Exceptional Children*, 69(2), 211–225.
- Herrera, A. W., Turner, C. D., Quenemoen, R. F., & Thurlow, M. L. (2015). *NCSC's age and grade-appropriate assessment of student learning* (NCSC Brief No. 6). Minneapolis, MN: University of Minnesota, National Center and State Collaborative.
- Huynh, H., & Barton, K. E. (2006). Performance of students with disabilities under regular and oral administrations of a high-stakes reading examination. *Applied Measurement in Education*, 19(1), 21–39.
- ITC (International Test Commission). (2018). ITC guidelines for the large-scale assessment of linguistically and culturally diverse populations (second edition). *International Journal of Testing*, 18, 101–134. <http://doi.org10.1080/15305058.2017.1398166>.
- Johnstone, C., Higgins, J., & Fedorchak, G. (2019). Assessment in an era of accessibility: Evaluating rules for scripting audio representation of test items. *British Journal of Educational Technology*, 50, 806–818.
- Kearns, J. F., Towels-Reeves, E., Kleinert, H. L., Kleinert, J. O., & Kleine-Kracht, M. (2011). Characteristics of and implications for students participating in alternate assessments based on alternate academic achievement standards. *Journal of Special Education*, 45(3), 3–14.
- Kettler, R. J. (2012). Testing accommodations: Theory and research to inform practice. *International Disability, Development, and Education*, 5(1), 53–66.
- Laitusis, C. C., Maneckshana, B., Monfils, L., & Ahlgrim-Delzell, L. (2014). Differential item functioning comparisons on a performance-based alternate assessment for students with severe cognitive impairments, autism and orthopedic impairments. *Journal of Applied Testing Technology*, 10(2), 1–33.
- Lazarus, S. S., & Thurlow, M. L. (2016). *2015-16 high school assessment accommodations policies: An analysis of ACT, SAT, PARCC, and Smarter Balanced* (NCEO Report No. 403). National Center on Educational Outcomes.
- Lee, A., Browder, D. M., Wakeman, S. Y., Quenemoen, R. F., & Thurlow, M. L. (2015, August). *AA-AAS: How do our students learn and show what they know?* (NCSC Brief No. 3). Minneapolis, MN: University of Minnesota, National Center and State Collaborative.

- Lewandowski, W., Wood, W., & Lambert, T. (2015). Private room as a test accommodation. *Assessment & Evaluation in Higher Education*, 40(2), 279–285.
- Li, H. (2014). The effects of read aloud accommodations for students with and without disabilities: A meta-analysis. *Educational Measurement Issues and Practice*, 33(3), 3–16.
- Lin, P. Y., & Lin, Y. C. (2014). Examining student factors in sources of setting accommodation DIF. *Educational and Psychological Measurement*, 74(5), 759–794.
- Lovett, B. J. (2010). Extended time testing accommodations for students with disabilities: Answers to five fundamental questions. *Review of Educational Research*, 80(4), 611–638.
- Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (pp. 13–103). Washington, DC: American Council on Education.
- Mislevy, R. J., Haertel, G., Cheng, B. H., Ructtinger, L., DeBarger, A., Murray, E., et al. (2013). A “conditional” sense of fairness in assessment. *Educational Research & Evaluation*, 19, 121–140. <https://doi.org/10.1080/13803611.2013.767614>.
- NCSC (National Center and State Collaborative). (2015). NCSC assessment policies. Retrieved from [www.ncscpartners.org/Media/Default/PDFs/Resources/Parents/NCSCAssessmentPolicies.pdf](http://www.ncscpartners.org/Media/Default/PDFs/Resources/Parents/NCSCAssessmentPolicies.pdf).
- NCEO (National Center on Education Outcomes). (2016). *Participation in AA-AAS*. Minneapolis, MN: University of Minnesota.
- NCME (National Council on Measurement in Education). (2019). *National Council on Measurement in Education Position Statement on the Use of College Admissions Test Scores as Academic Indicators in State Accountability Systems*. Available at [https://higherlogicdownload.s3.amazonaws.com/NCME/c53581e4-9882-4137-987b-4475f6cb502a/UploadedImages/Documents/Admission\\_Statement\\_06-16-19.pdf](https://higherlogicdownload.s3.amazonaws.com/NCME/c53581e4-9882-4137-987b-4475f6cb502a/UploadedImages/Documents/Admission_Statement_06-16-19.pdf).
- NRC (National Research Council). (2004). *Keeping score for all: The effects of inclusion and accommodation policies on large-scale educational assessments* (J. A. Koenig & L. F. Bachman, Eds.). Washington, DC: The National Academies Press.
- PARCC (Partnership for Assessment of Readiness for College and Careers). (2016). *PARCC accessibility features and accommodations manual 2016–2017* (5th ed.). Washington, DC: Author.
- Searcy, C. A., Dowd, K. W., Hughes, K. G., Baldwin, S., & Pigg, T. (2015). Association of MCAT scores obtained with standard vs extra administration time with medical school admission, medical student performance, and time to graduation. *Journal of the American Medical Association*, 313(22), 2253–2262. <http://doi.org/10.1001/jama.2015.5511>.
- Shifrer, D., & Fish, R. (2019). Contextual reliability in the designation of cognitive health conditions among U.S. children. *Society and Mental Health*. <http://doi.org/10.1177/2156869319847243>.
- Sireci, S. G. (2005). Unlabeling the disabled: A perspective on flagging scores from accommodated test administrations. *Educational Researcher*, 34, 3–12.
- Sireci, S. G., Banda, E., & Wells, C. S. (2018). Promoting valid assessment of students with disabilities and English learners. In S. N. Elliott, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), *Handbook of accessible instruction and testing practices: Issues, innovations, and application* (pp. 231–246). New York: Springer.
- Sireci, S. G., & Faulkner-Bond, M. (2015). Promoting validity in the assessment of English learners. *Review of Research in Education*, 39(1), 215–252.
- Sireci, S. G., & Gandara, M. F. (2016). Testing in educational and developmental settings. In F. Leong et al. (Eds.), *International Test Commission handbook of testing and assessment* (pp. 187–202). Oxford, UK: Oxford University Press.
- Sireci, S. G., Scarpati, S., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research*, 75, 457–490.
- Sireci, S. G., Wells, C., & Hu, H. (2014, April). *Using internal structure validity evidence to evaluate test accommodations*. Paper presented at the annual meeting of the National Council on Measurement in Education, Philadelphia.
- Smarter Balanced. (2016). *Smarter Balanced Assessment Consortium: Usability, accessibility and accommodations guidelines*. Retrieved October 10, 2019, from <http://www.smarterbalanced.org/wp-content/uploads/2015/09/Usability-Accessibility-Accommodations-Guidelines.pdf>.
- Thompson, S., Blount, A., & Thurlow, M. (2002). *A summary of research on the effects of test accommodations: 1999 through 2001* (Technical Report No. 34). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved January 2003 from <http://education.umn.edu/NCEO/OnlinePubs/Technical34.htm>.

- Thurlow, M. L., Lazarus, S. S., Christensen, L. L., & Shyyan, V. (2016). *Principles and characteristics of inclusive assessment systems in a changing assessment landscape* (NCEO Report No. 400). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Wells-Moreaux, S., Bechard, S., & Karvonen, M. (2015). *Accessibility manual for the dynamic learning maps alternate assessment, 2015–2016*. Lawrence, KS: The University of Kansas Center for Educational Testing and Evaluation.
- Zebehazy, K. T., Zigmond, N., & Zimmerman, G. J. (2012). Performance measurement and accommodation: Students with visual impairments on Pennsylvania's alternate assessment. *Journal of Visual Impairment & Blindness*, 106(1), 17–30.
- Zerkel, P. A., & Weathers, J. M. (2016). K–12 students eligible solely under section 504: Updated national incidence data. *Journal of Disability Policy Studies*, 27(2), 67–75. <http://doi.org/10.1177/1044207315626115>.



# Comparability in Multilingual and Multicultural Assessment Contexts

Kadriye Ercikan, *Educational Testing Service/University of British Columbia*  
Han-Hui Por, *Educational Testing Service*<sup>1</sup>

## CONTENTS

INTRODUCTION .....	205
SOCIOCULTURAL AND LANGUAGE CONSIDERATIONS IN ASSESSMENT .....	207
Impact of Sociocultural and Language Factors on Score Interpretation and Use. . .	208
Policies Addressing Language Diversity .....	209
MEASUREMENT EQUIVALENCE .....	211
Degrees of Measurement Incomparability .....	212
Sources of Measurement Incomparability .....	213
Limitations of DIF Methodology .....	214
GUIDELINES AND CONSIDERATIONS FOR MULTIPLE LANGUAGE	
VERSIONS OF ASSESSMENTS .....	214
AERA, APA, and NCME Standards. ....	215
ITC Test Adaptation Guidelines. ....	215
Assessment Development and Adaptation Processes. ....	216
Creating Comparable Scores. ....	217
NEXT-GENERATION ASSESSMENTS .....	218
CONCLUSION .....	220
REFERENCES .....	220

## INTRODUCTION

A basic tenet of the validity of inferences from assessments is that scores reflect the underlying knowledge and abilities that the test is designed to measure, and the score meaning is consistent for individuals from different language and sociocultural backgrounds. The validity of interpretation of performance on assessments in multicultural and multilingual contexts is critically tied to whether (1) the assessments are tapping the

---

<sup>1</sup> The authors are grateful to Randy Bennett, Maria Elena Oliveri, and Donald Powers for reviewing an earlier draft of this chapter; to Guadalupe Valdés and Christian Faltis for their feedback; and to Emily Pearce for editorial support.



knowledge and skills we are interested in assessing, (2) the constructs being assessed are comparable for different sociocultural groups, and (3) the scores are comparable across languages and cultures (Ercikan & Lyons-Thomas, 2013). These criteria for score comparability are at the heart of fairness in the interpretation and use of assessment results; they require us, as assessment developers, users, and specialists, to examine and verify what constructs the assessments are targeting and whether they are assessing the same construct with the same psychometric properties for different groups.

Ensuring that assessments provide consistent score meaning is crucial when students have different language and sociocultural backgrounds. In international assessments of learning outcomes, such as the Trends in International Mathematics and Science Study (TIMSS), the Progress in International Reading Literacy Study (PIRLS), and the Programme for International Student Assessment (PISA), consistency of score meaning across countries, languages, and cultural groups is central to making accurate and meaningful inferences. Chapter 5 (Comparability Across Different Assessment Systems) elaborates on the validity of the comparisons made across different assessment systems. In addition to international assessments, the issue of consistent score meaning is also a concern for countries with populations from diverse language and sociocultural backgrounds such as in countries with large immigrant populations. In the United States, students have varied sociocultural and language backgrounds, with large proportions speaking a language at home and in their community that differs from the language used in school. The recognition of diversity and its implications for validity and fairness led states to develop assessments in multiple languages and to provide language tools and accommodations to assess students' performance. For example, in New York State, mathematics assessments are adapted into such student home and community languages as Spanish, Traditional Chinese, Haitian, Korean, and Russian (Tabaku, Carbuccia-Abbott, & Saavedra, 2018).

The purpose of this chapter is to highlight the complexity of comparability issues when assessments are administered in multiple languages to students from diverse backgrounds, to describe research on the comparability of assessments and scores, and to discuss guidelines and processes in optimizing comparability of multiple language versions of assessments. The first section describes the sociocultural and language diversity in the United States and countries around the world and discusses the impact of such diversity on interpretation and use of assessment results. The next section introduces the concepts of measurement equivalence and the methodologies used in examining measurement equivalence and score comparability. The third section describes the guidelines for optimizing score comparability across adapted versions of assessments and provides recommendations to create comparable scores. The final section addresses score comparability challenges and potential solutions for the next-generation assessments that involve administration in digital environments.

Throughout this chapter, we distinguish between *adaptation* and *translation*, with the latter term used to denote creating different language versions with a focus on linguistic equivalence. *Adaptation*, however, refers to the broader process of creating language versions that may include changes made to create greater cultural relevance in addition to linguistic equivalence. The term *adaptation* is preferred in assessment contexts because the task goes beyond the literal translation of the assessment content and more accurately reflects the process that is expected to lead to a greater validity

of the assessment for the targeted populations. Furthermore, *measurement/score comparability* and *measurement equivalence* are used interchangeably to broadly define the comparability aspects of assessments that include comparability of score interpretation as well as statistical notions of measurement equivalence.

### SOCIOCULTURAL AND LANGUAGE CONSIDERATIONS IN ASSESSMENT

Sociocultural and language diversity is a reality shared by all countries around the world. In the U.S. context, information about languages spoken has been obtained from Census respondents since 1980. According to the Census data, the estimated percentage of people speaking *only* English at home has steadily fallen, declining from 89.1 percent in 1980 to 78.2 percent in 2017. Other widely used languages include Spanish (41.0 million), Chinese (3.5 million), Tagalog (1.7 million), Vietnamese (1.5 million), Arabic (1.2 million), French (1.2 million), and Korean (1.1 million) (U.S. Census Bureau, 2017a). In the 2017 Census, 80 percent of the school-age children who spoke a different language at home and in their communities also spoke English “very well”<sup>2</sup> (U.S. Census Bureau, 2017b). The extent of language diversity raises the issue of how best to assess students in ways that lead to valid and fair interpretation and use of assessments. In particular, questions arise regarding what language students should be tested in: the home/community language versus the language of schooling? Or should both languages be included in the assessment (López, Turkan, & Guzmán-Orth, 2017)? What kinds of language accommodation tools should be provided to students, and how can score comparability be established across multiple language versions of assessments?

The inherent differences between languages can make test adaptation a challenging task. Adapted versions of assessments must reflect equivalent meaning, format, relevance, intrinsic interest, and familiarity of the item content (Ercikan, 1998, 2003; Hambleton, Merenda, & Spielberger, 2005). Languages vary in the frequency of word use and word difficulty. Moreover, grammatical forms in one language may not have equivalent forms in other languages, or may possibly have many of them. There is also the difficulty of adapting syntactical style from one language to another. Languages may also differ in form (alphabet versus character based) and direction of scribe (left-to-right, right-to-left, or top-to-bottom). Some languages, such as German and French, also require much more text to convey the same intended meaning compared to English.

Sociocultural factors must also be taken into account when developing assessments and in interpretation and use of scores (Geisinger, 1994; McQueen & Mendelovits, 2003; Wu & Ercikan, 2006). Learners from diverse sociocultural backgrounds have different experiences with schooling and learning. For example, cultural norms can affect learning and world views, including how success is perceived. How students are taught, and how achievement is defined in the educational system, generally reflects the perspectives of the mainstream society and not necessarily those of all its cultural groups. For example, Yup’ik children in rural Alaska learn critical community practices, such as fishing and navigation, from observing and participating in these activities with experienced adults. Because verbal interactions are part of this key learning process,

---

<sup>2</sup> The available response options to the survey item “How well does this person speak English?” were “Very well,” “Well,” “Not well,” and “Not at all.”

a school system that expects passive listening with little interactions may put these students at a disadvantage (Lipka & McCarty, 1994).

Diverse sociocultural backgrounds also include the students' socioeconomic status (SES), which affects their experiences with schooling and learning, which in turn affect how the students interact with assessments. Research has provided ample evidence that higher SES is associated with higher achievement (Berliner, 2012; Lee & Burkam, 2002; Perry & McConney, 2010; Sirin, 2005; Tate, 1997), such as in reading (Grover & Ercikan, 2017; Silva, Verhoeven, & van Leeuwe, 2011). The positive association between SES and achievement is consistent with its relations to other life outcomes, including health status (McEniry, Samper-Ternent, Flórez, Pardo, & Cano-Gutierrez, 2019). High-SES students, as a group, attend schools that provide them with more resources, such as a higher teacher-to-student ratio (Shifrer & Fish, 2019). Indeed, SES has been shown to affect students' access to academic preparations (Carnevale & Rose, 2003), the identification and availability of aids for students with learning disabilities (Elder, Figlio, Imberman, & Persico, 2019), opportunity to learn (Bachman, Votruba-Drzal, El Nokali, & Castle Heatly, 2015; Blömeke, Suhl, Kaiser, & Döhrmann, 2012), and eligibility for test accommodations (Zirkel & Weathers, 2016). SES is also related to one's eligibility for unwarranted time extensions. Quealy and Shapiro (2019) reported that white, middle-class students were much more likely than students of other races and SES backgrounds to receive section 504 plan provisions, unfairly allowing students twice as much time to complete the New York specialized high school entrance examinations. The interaction between assessments and students' socioeconomic backgrounds suggests that taking these contexts into account in developing and interpreting assessment results can prevent further disadvantaging teachers and students from low-SES schooling contexts. A more detailed discussion of the disparities in access to test accommodations is found in Chapter 7, *Comparability When Assessing Individuals with Disabilities*.

Issues of displacements should also be taken into consideration when interpreting and using assessment results. Studies have shown that students who are displaced—whether due to school closures (Kirshner, Gaertner, & Pozzoboni, 2010) or natural disasters (e.g., Hurricane Katrina; Ward, Shelley, Kaase, & Pane, 2008) or because they are refugees (Gahungu, Gahungu, & Luseno, 2011)—showed declines in academic performance. In these instances, scores from these students may not truly reflect their actual ability, and their scores have to be interpreted with those considerations.

### **Impact of Sociocultural and Language Factors on Score Interpretation and Use**

An important issue often overlooked is that students' performances on assessments are the outcomes of complex interactions between the students' experiences (such as their language and sociocultural backgrounds) and the knowledge and skills targeted by the assessment and other properties of the assessment. When students interact with or participate in assessments, multiple psychological, social, and cognitive factors are at play, including their home/community language; their familiarity with contexts, objects, words, and how students relate to them; how they function in an assessment context; and their anxiety levels, among many other affective and conative variables (Snow, 1993). Solano-Flores and Nelson-Barber (2001) added other

sociocultural influences prevalent in cultural groups, such as communication patterns and socioeconomic conditions.

While responses to assessment questions may reflect what students know and can do, their cultural and language practices outside of school also affect how they interpret assessment questions and formulate their responses. Ercikan, Roth, Simon, Sandilands, and Lyons-Thomas (2014) noted that the nature and frequency of access to white mainstream cultural practices outside of school contribute to students interpreting items differently. Parents who do not understand the purpose or nature of formal assessments or who have not been invited into mainstream school cultural practices may also contribute to students' confusion and to an increase in test anxiety. Likewise, teachers who are unable to communicate well with the parents or all of the children in their classes due to language differences may also contribute to students' confusion and test anxiety. Furthermore, differences in sociocultural norms can lead to differences in how students engage with assessments and what responses they provide to test items (Solano-Flores, Lara, Sexton, & Navarrete, 2001). For example, students from some cultural groups may have been socialized to not provide lengthy or elaborate answers to interviewers who are considered to be of higher status or maturity, or they may hold back when indicating confidence and success level or when being requested to disclose personal information. Possible sociocultural differences may also be found in motivation, experience with psychological assessments, and speed of responding (Talento-Miller, Guo, & Han, 2013).

Another important consideration that can affect the students' assessment performance is their access to opportunities to learn and engage with similar kinds of assessments (Ercikan, Roth, & Asil, 2015). The opportunity to learn the curricular content which is subsequently assessed, develop test-taking strategies, and become familiar with the assessment technology, as in the case of digitally-based assessments, can all contribute to the students' ability to engage with the assessment. Hambleton (2005) highlighted item format as a potential threat to score comparability, which suggests that currently novel item types (e.g., hot zone selection, drag and drop) should be used with caution when assessing students with nonwhite and nonmainstream language and sociocultural backgrounds to minimize introducing construct-irrelevant demands in assessment questions. Access to test preparation classes or test time extension can further contribute to measurement incomparability. Given that the scores should reflect primarily the competencies being assessed (Messick, 1989, 1995), the interpretations of scores across different language and sociocultural groups should account for the differences in access to, participation in, and benefit from learning opportunities.

### **Policies Addressing Language Diversity**

The recognition of language and cultural diversity among the student population in the United States and its implications for validity and fairness moved states to provide language tools and accommodations to assess students' performance. These language tools and accommodations are intended to optimize student performance and minimize the impact of language proficiency on performance on assessments that are not intended to assess language proficiency. Policies dealing with language diversity within the United States vary widely among jurisdictions (Tabaku et al., 2018). The

Smarter Balanced Assessment Consortium provides language support for students in the form of glossaries and translations of test directions and items in several languages commonly spoken as home and community languages. Standard language glossaries are available in Spanish, Arabic, Cantonese, Mandarin, Filipino (Ilokano and Tagalog), Korean, Punjabi, Russian, Ukrainian, and Vietnamese (Smarter Balanced, n.d.). However, in each state students' access to such tools and support depends on state laws and regulations.

Currently, some states allow alternative standard language versions of some assessments (New York State Education Department, 2016; Ohio Department of Education, 2018, 2019; Oregon Department of Education, 2019). In New York, alternative languages include Spanish, Chinese, Haitian Creole, Russian, Polish, Korean, Bengali, Arabic, Urdu, Vietnamese, Amharic, Portuguese, and several others. Other states translate test directions, but not the assessment itself, into commonly spoken home/community languages (e.g., South Carolina Department of Education, n.d.; State of New Jersey Department of Education, 2019). More frequently, students who need language support are provided with word-to-word or translation dictionaries, which give standard language counterparts for specific terms but not definition, use, or explanation (Florida Department of Education, n.d.; Ohio Department of Education, 2018, 2019; South Carolina Department of Education, n.d.; State of New Jersey Department of Education, 2019). Chapter 6, *Comparability When Assessing English Learner Students*, provides more details on accommodations designed to help students from non-English and/or bilingual backgrounds.

Growing language and sociocultural diversity is not unique to the United States. Linguistic and sociocultural diversity exists to varying degrees throughout the world, and the recognition and treatment of such diversity also vary. South Africa, which has 11 official languages, administers assessments in these languages to students who come from backgrounds that include dozens of other languages spoken in the community. In Canada, which has two official languages, English and French, assessments are given in the two official languages, and students are assessed in the language of instruction (Ercikan, Oliveri, & Sandilands, 2013).

Placing value on language and sociocultural diversity necessitates policies to ensure that, regardless of background, students have the same opportunities to demonstrate their knowledge, skills, and competencies on assessments. This may necessitate exempting students whose English language proficiency has not advanced enough to allow them to demonstrate their knowledge, skills, and competencies using assessments conducted entirely in English. Only when bilingual students have developed the required level of language proficiency should they be tested in English and be provided the tools and accommodations to support their performance on assessments. Another possibility is to provide education and assessment in the students' home/community language until they have developed enough English proficiency to fully participate in and benefit from an education in English. In such cases, students will be administered adapted versions of assessments in their home language.



## MEASUREMENT EQUIVALENCE

Developing assessments that capture the intended set of constructs for a language and sociocultural group requires extensive research to provide insights on how the construct is operationalized and developed in different contexts and empirical evidence that the assessment captures the intended constructs. The challenges are multiplied in multilingual and multicultural assessment contexts when the targeted constructs differ across and within cultural groups and social contexts, or the assessments are adapted to different languages. The appropriate interpretation of scores and comparisons for students from different sociocultural contexts and language backgrounds requires establishing empirical evidence of measurement equivalence for the considered groups. Measurement equivalence includes (1) *construct equivalence*, (2) *test equivalence*, and (3) *equivalence of testing conditions* (Ercikan & Lyons-Thomas, 2013; also see Chapter 5, Comparability Across Different Assessment Systems).

*Construct equivalence* is defined as the equivalence of meaning of the construct in terms of its theoretical definition, the way it is operationalized, and the way it is developed for the cultures in which the assessment will be administered. In addition to demonstrating similar psychometric properties, the evidence for whether a construct (e.g., reading proficiency) is conceptualized the same way across language and sociocultural groups also needs to be grounded in empirical research based on the considered language and sociocultural groups.

*Test equivalence* refers to the equivalence of test content, and the language and sociocultural equivalence at the item and the overall test levels. This includes the equivalence of text, graphics, formatting, language meaning, language demands, and cues for responding to test questions, and the sociocultural relevance in the words and contexts provided in the items.

*Equivalence of testing conditions* refers to the equivalence of test administration conditions such as the communication between test administrator and examinees, which includes test directions, instructions, and training sessions for the test administrators. It includes whether the different language versions of tests were administered in an identical fashion, whether the test format was equally appropriate in each language version, whether the speed of response was not more of a factor in one language than the other, and whether other response styles such as acquiescence, tendency to guess, and social desirability did not vary significantly across groups (Hambleton, 2005; Hambleton & Patsula, 1999). In addition, in a different language or sociocultural setting, test administrators should be drawn from the local language communities; be familiar with the culture, language, and local dialects; have adequate test administration skills; and know the importance of following standardized procedures associated with the assessment. Broader testing conditions should also be considered in interpreting scores of students from diverse cultural and language backgrounds. These conditions may include societal context for testing (such as the emphasis given to testing), which may affect how students perceive the testing situation and the role of testing as well as students' motivation to perform and how they engage with the assessment.

In addition to measurement equivalence, score comparability requires measurement unit or scalar equivalence, which refers to whether units on the score scales based on different assessment versions have equivalent units. Even when measurement equivalence requirements are met, in order to compare performance levels of students from



different language and sociocultural backgrounds, or test forms taken in different languages, scalar equivalence is required. For example, a score difference of 10 score points on one scale based on the source-language version of the assessment can only be considered equivalent to 10 score points on a scale based on the non-English language version when the scores are on the same scale and using the same measurement units.

While measurement equivalence of items across subgroups is ideal, true equivalence is difficult to achieve in practice. In reality, variations across assessment versions or across language and sociocultural groups are inevitable. For instance, items may exhibit varying degrees of differential item functioning (DIF). The classification scheme developed at the Educational Testing Service assumes functional equivalence if the items exhibiting DIF do not exceed statistical significance threshold values (Dorans & Holland, 1992). Similarly, empirical evaluations of equivalence require determining the levels of difference for different language and sociocultural groups and establishing what levels of differences may be tolerated without compromising the comparability of scores for these groups.

### **Degrees of Measurement Incomparability**

The growing interest in international assessments and comparisons, and a recognition of the complexity of score comparability across language and culture groups, have led to extensive research on the comparability of assessments and scores across languages, the procedures that optimize the performance of students from different language backgrounds, and the development of guidelines for test adaptation. Research on item-level comparability focuses on “unexpected” performance differences for examinee subgroups that are matched in terms of their overall ability or performance on the test. Regardless of group membership (e.g., based on gender, ethnicity, socioeconomic, or language backgrounds), students with the same ability level should have the same likelihood of receiving the same test score. The great majority of this research has focused on using DIF methods to examine item equivalence (e.g., Dorans & Kulick, 1986; Ercikan, 1998; Ercikan & Lyons-Thomas, 2013; Gierl, Rogers, & Klinger, 1999; Oliveri, Olson, Ercikan, & Zumbo, 2012; Padilla, Benítez, & Castillo, 2013; Sireci & Allalouf, 2003) and measurement equivalence based on test data factor structure (e.g., Ercikan & Koh, 2005; Güzel & Berberoglu, 2005; Sireci, Patsula, & Hambleton, 2005; Zumbo, 2003). More comprehensive discussions of DIF detection methods can be found in Holland and Wainer (2012) and Chapter 3, *Comparability of Aggregated Group Scores on the “Same Test.”*

Research using DIF methodology has demonstrated extensive incomparability between language versions of assessments within and across countries. Studies conducted in Canada comparing the French and English versions of large-scale assessments found that 18 to 60 percent of items functioned differently for the two language groups (Ercikan, Gierl, McCreith, Puhan, & Koh, 2004), pointing to high levels of incomparability at the item level. Research on international assessments identified high levels of incomparability between different language versions of items and tests (Byrne & van de Vijver, 2010; Ercikan, 1998; Ercikan & Koh, 2005; Ercikan & Lyons-Thomas, 2013; Ercikan & McCreith, 2002; Gierl, 2000; Gierl et al., 1999; Grisay, 2003; Hambleton et al., 2005; Marotta, Tramonte, & Willms, 2015; Oliveri & Ercikan, 2011; Solano-Flores, Backhoff,

& Contreras-Niño, 2009). For example, Ercikan and Koh (2005) compared English and French versions of TIMSS administered in the United States and France and identified that 79 percent of the science items functioned differentially between the two countries and language groups. These findings highlight the importance of the quality of the adaptation process on the validity of measurement and the comparability of scores, and the importance of establishing measurement comparability of assessments across languages.

Some research evidence demonstrates that measurement incomparability identified at item levels does not necessarily lead to observable scale-level differences. Previous studies using various methods such as exploratory factor analyses (Arim & Ercikan, 2005), confirmatory factor analyses (Ercikan & Koh, 2005; Oliveri et al., 2012; Zumbo, 2003), and test characteristic curve (TCC) comparisons (Ercikan & Gonzalez, 2008) identified little to no differential test functioning (DTF) despite large proportions of differences identified at the item level. For example, Ercikan and Gonzalez (2008), using item response theory-based TCCs, found small score-scale differences between different language versions of PIRLS assessments despite the presence of large percentages of DIF items. Also, as shown by a study conducted by Zumbo (2003) using confirmatory factor analysis, a similarly negligible DTF was seen even when the test contained large amounts of high-level DIF against a group. The inconsistency between item- and scale-level differences in measurement equivalences point to the importance of examining measurement comparability both at the item level and the scale level. Furthermore, one of the purposes of DIF analysis in multilingual assessments is to help identify possible differences created by the test adaptation process. Identifying such possible adaptation problems before assessments are finalized can provide opportunities for correcting adaptation problems and enhance comparability.

### **Sources of Measurement Incomparability**

Even though one purpose of DIF analyses is to identify adaptation problems, the sources of statistical differences identified by DIF are often difficult to pinpoint (Ercikan, 2002; Haberman & Dorans, 2011; Hambleton et al., 2005). In particular, DIF found in items between language versions of an assessment does not always indicate problems in test adaptation. Sources of DIF are often explored using bilingual expert reviews and think-aloud protocols with students from non-English language groups (Ercikan et al., 2004). For example, Ercikan and McCreith (2002) examined sources of DIF using bilingual experts for comparing items in the English and French versions of TIMSS and demonstrated that in some booklets as few as 36 percent of the cases evidencing DIF were due to adaptation problems. Even when differences are identified by bilingual experts, these sources of DIF must be treated as hypotheses, as other research has shown that adaptation differences do not necessarily lead to differences in how students read, interpret, and solve test items in different languages (Ercikan et al., 2004).

Researchers have recognized the complexity of identifying potential sources of score incomparability in assessing students from diverse backgrounds (Ercikan, 2002; Hambleton et al., 2005). Psychometric differences between language versions can be due to multiple factors, including sociocultural and curricular differences between groups (Ercikan et al., 2004; Solano-Flores & Nelson-Barber, 2001; Solano-Flores, Trumbull,

& Nelson-Barber, 2002), educational policies and standards, values, and motivation to take the assessment (Arffman, 2010; Gee, 2013; Greenfield, 1997; John-Steiner & Mahn, 1996). Cultural differences can influence intrinsic interest in, familiarity with, and the interpretation of item content. In addition, word meaning is created through social and cultural interactions (Campbell, 2003; Derrida, 1998; Greenfield, 1997), and social context and cultural experiences are expected to affect interpretations of words, which in turn may affect the trajectory of thought processes and ultimately responses to assessment questions (Ercikan & Lyons-Thomas, 2013; Ercikan & Roth, 2006; Roth, 2009, 2010; Solano-Flores, 2006). Even within the same language context, such as in the United States or Canada, students from some sociocultural groups speak structurally and semantically different varieties of English. Examples of these include some but not all indigenous students, African-American students, Mexican-American students, and students from nonmainstream-SES backgrounds. For some but not all of these students, written standard English is an added difficulty (Roth & Harama, 2000), thereby creating linguistic incomparability for students from different sociocultural backgrounds.

### **Limitations of DIF Methodology**

Research results increasingly point to the limitations of methodologies used in examining measurement comparability when the diverse sociocultural context is not taken into account. This research indicates that neglecting the within-group heterogeneity for DIF methodology has validity implications (Grover & Ercikan, 2017). Using simulated data, Oliveri, Ercikan, and Zumbo (2014) varied the heterogeneity within the focal groups from 0 to 80 percent and found that, as heterogeneity increased, the rates of correct DIF detection decreased. Ercikan and colleagues (2014) conducted DIF analyses on heterogeneous linguistic groups created using the information on the language of instruction, dominant language setting, and home/community language. They demonstrated that English and French language learners are heterogeneous groups, and that the DIF results do not necessarily apply to all members of a given language group in the same way, suggesting that issues other than test adaptations also play a role.

Researchers have attempted to account for heterogeneity by crossing two manifest groups (e.g., gender and ethnicity) to create more specific groups for DIF analysis. DIF analyses conducted at the subgroup level allow the researcher to determine whether and to what extent DIF items detected at the population level are the same as the DIF items detected at various subpopulation levels. This has been referred to as a “melting pot” DIF (Dorans & Holland, 1992) or DIF dissection approach (Zhang, Dorans, & Matthew-López, 2005). Ercikan and Oliveri (2013) also proposed a two-step approach in conducting DIF on heterogeneous samples in which latent class analysis is conducted within groups of the manifest variable of interest (i.e., males and females), as opposed to on the whole sample, which is typically done.

### **GUIDELINES AND CONSIDERATIONS FOR MULTIPLE LANGUAGE VERSIONS OF ASSESSMENTS**

Two sets of guidelines dedicate significant attention to test adaptation and evaluation of the quality of test adaptations. These are (1) the standards developed by the American Educational Research Association (AERA), the American Psychological

Association (APA), and the National Council on Measurement in Education (NCME), and (2) the International Test Commission (ITC) guidelines. Relevant sections of these guidelines for multilingual and multicultural versions of assessments are summarized below.

### AERA, APA, and NCME Standards

In the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999, 2014), four standards are particularly relevant to test adaptation:

- **Standard 9.4** highlights the need for assessment developers to explain and provide justification for linguistic modifications that they deem to be appropriate in specific situations. These modifications should be taken into account in score interpretations.
- **Standard 9.5** recommends that if there is evidence that scores are not comparable across multiple versions of assessments, additional information should be provided to help users interpret assessment scores correctly.
- **Standard 9.7** calls attention to the need to describe the approaches used in establishing the adequacy of adaptation, and for empirical and logical evidence to be provided for score reliability and the validity of the inferences based on the target assessment for all linguistic groups. For example, if an assessment adapted into Spanish is meant to be used with Mexican, Cuban, Spanish, and other Spanish-speaking subgroups, it is the responsibility of the assessment developer to provide independent reliability and validity evidence for each of those subgroups.
- **Standard 9.9** recommends that assessment developers provide evidence of the comparability of different language versions of an assessment. For example, the assessment developer should present evidence that the same construct is being measured in both assessments.

The Standards also advise against using back-translation, which involves comparisons of the source version with the target-to-source translation, as the sole method for verifying linguistic comparability. The comparability of the source- and back-translated target versions is not sufficient to establish that the two language versions have the same meaning and provide similar information to students. The use of interpreters or translators who are not familiar with proper testing procedures or purposes of testing may make inadequate translation and adaptation of the assessment and provide inappropriate test administration.

### ITC Test Adaptation Guidelines

The second set of guidelines was developed by the ITC (Hambleton, 2005; ITC, 2018). The ITC emphasizes these guidelines as being instrumental in conducting and evaluating the adaptation or the simultaneous development of assessments for use with different populations. The 18 guidelines, along with suggestions for practice, were organized around six broad topics: pre-condition, test development, confirmation/empirical analyses, administration, score scales and interpretation, and documentation.

The guidelines in the section titled “Pre-Condition” highlight the decisions that are made before the translation/adaptation process begins. The second section, “Test Development,” focuses on the process of adapting an assessment that includes items, test instructions, and scoring rubrics. The third section, “Confirmation,” contains guidelines on documenting empirical evidence addressing score equivalence, as well as reliability and validity of the assessment in multiple languages and cultures. The section “Administration” pertains to the preparation of materials and instructions to minimize language and sociocultural issues in the administration process. The fifth section, “Score Scales and Interpretation,” discusses score comparisons. The final section, “Documentation,” contains guidelines detailing the technical aspects of test adaptations and the appropriate use of the test scores.

### Assessment Development and Adaptation Processes

The quality of adaptation is optimized when assessments in the source language are developed with the test adaptation goal in mind. Brislin, Lonner, and Thorndike (1973) suggested using short, simple sentences of fewer than 16 words, employing the active rather than the passive voice, repeating nouns instead of using pronouns, and using specific rather than general terms (e.g., “cows, chickens, and pigs” rather than “livestock”). For an assessment to be adaptable, the source assessment should also avoid the language structures that are unlikely to have equivalents in other languages, such as metaphors or colloquialisms, certain modals (e.g., verb forms with “could” or “would”), adverbs and prepositions telling “where” or “when” (e.g., frequent, beyond, or upper), possessive forms (e.g., mine or Tim’s), probabilistic words (e.g., “probably” and “frequently”), and sentences with two different verbs if the verbs suggest different actions.

The most commonly used method for creating adapted versions of tests is *successive* test adaptation, where the assessment is developed in a source language and one or more bilingual translators adapt the assessment to the target language and culture. *Simultaneous* and *concurrent/parallel* assessment development are alternatives to the traditional approach of translating assessments created in a single source language, typically English in the North American context. In *simultaneous* assessment development, the emphasis is on the use of a multidisciplinary committee of experts in the languages, psychometrics, and the content domain for developing items (Tanzer & Sim, 1999). Test items are developed by *bilingual item writers* in one language and are immediately adapted into the other language. The *concurrent/parallel* assessment development model (Solano-Flores et al., 2002) utilizes *shells* or *templates*, which define item structure and the cognitive demands of each item. Using these templates for item development, the language groups work jointly in all stages of the assessment development process. The different language versions are developed by experts from each language group based on a common assessment blueprint. Using this approach, each assessment originates in the language it is targeted for and is developed by content experts from the particular sociocultural group, except for a small portion of the assessment that is adapted from the source language for linking purposes. Different development processes may have trade-offs between *comparability* and *cultural authenticity* of adapted assessments. While concurrent/parallel development prioritizes cultural authenticity, successive



development prioritizes comparability, and simultaneous assessment developments target a compromise between comparability and cultural authenticity (Ercikan & Lyons-Thomas, 2013).

### Creating Comparable Scores

There has been significant research on creating comparable score scales for multilingual versions of assessments (Cook, 2006; Cook & Schmitt-Cascallar, 2005; Sireci, 1997, 2005). This research indicates that, in the absence of sufficient evidence for measurement equivalence across groups, score scales should be based on separate language/country calibrations and comparability should be established through a linking procedure. Cook and Schmitt-Cascallar (2005) provide an overview of score comparability and describe four methods of linking scores on assessments given in different languages.

Currently, some of the most well-known assessments with multiple language versions include international assessments such as the TIMSS, the PISA, and the PIRLS. In these assessments, a single international score scale allows for comparisons of performances across countries and language groups. The single score scale is computed with item parameters calibrated on an international sample that consists of a randomly selected subsample from countries that take the assessments in different languages. In some assessments, country-specific parameters are used when the international item parameters do not fit the scale well because of some level of measurement incomparability.

In establishing comparability of scores across language versions of assessments, consideration of comparability must start from the conceptualization of the assessment. Ercikan and Lyons-Thomas (2013) identify seven key steps in developing and adapting assessments for use in different languages and cultures:

1. **Examine the equivalence of the constructs.** This step requires examining and comparing the construct definitions in the source and other cultures and languages. It may involve a review by cultural and language expert groups who can evaluate the appropriateness of the construct definitions and identify aspects of the construct that may be different for the two language and cultural groups.
2. **Select a test adaptation and development method.** The next step involves deciding on the approach to developing multiple versions of assessments. If a source version already exists, successive test adaptation may need to be used. If, however, there is an opportunity to build multiple language versions from the beginning, parallel or simultaneous development may be employed.
3. **Perform the adaptation of the test or measure.** There are several factors that will affect the quality of the adaptation. First is the linguistic features of the source version that might affect translatability. These include using short sentences, repeating nouns instead of pronouns, and avoiding metaphors and a passive voice in developing assessments. Other factors include language background and proficiencies of translators in the relevant languages, such as whether translators are fluent in both languages and knowledgeable about both source and target cultures, and they have clear understanding of the construct being assessed.



4. **Evaluate the language equivalence between the two assessment versions.** A necessary step in developing multiple language versions of assessments is an evaluation of equivalence by bilingual experts. Reviews of equivalence can determine differences in language, content, format, and other aspects of items in the comparison languages. Insights from such reviews can help inform revisions of adaptations to optimize comparability.
5. **Document changes made in the adaptation process.** Documenting changes and the rationale for these changes between the language versions of assessments is critical for informing test users for potential impact on comparability.
6. **Conduct a field test study to examine measurement equivalence.** Establishing measurement equivalence requires empirical evidence to support such an evaluation. Field test data can be used to examine the reliability and validity of both language versions of assessments, as well as measurement equivalence using classical test theory-based analyses, factor analyses, DIF analyses, and comparisons of TCC curves. Additional follow-up studies can include a second round of expert reviews and cognitive analyses to provide further support for comparability of the language versions of assessments.
7. **Conduct linking studies.** In order to create comparable scales with measurement unit equivalence, a linking study is needed once measurement equivalence has been established.

### NEXT-GENERATION ASSESSMENTS

As assessments increasingly include multiple modes of administration—using both paper and digital delivery, and sometimes delivered on multiple types of digital devices within the same test administration—as well as increasingly complex forms of interactivity, assessment developers need to consider ways student backgrounds may affect how students engage with assessments. The technology-enhanced environments provide growing opportunities for interactivity between the student and the assessments, and usually involve multiple modes of engaging with the assessment. To understand the questions, students may be required to read excerpts, listen to audio segments, view video clips, or manipulate diagrams with their mouse. They then respond by typing text, speaking, drawing diagrams, plotting graphs, or dragging items. While the first generation of computer-based assessments typically mimics its paper-based counterparts in using text-based items, advancing technology has made new item formats possible. An example is the integration of videos in assessment items, which allows for the assessment stimulus to be delivered through an acted-out scenario. Videos (or animations) are useful when a detailed description can be too lengthy and the video can more effectively demonstrate and elaborate on what is being described. In assessments of language proficiency, students can be asked to verbally describe the interactions they see in a video, which may approximate how language is used in day-to-day life. Mechanical reasoning can also be tested when students are presented with videos of machine components and given verbal or written questions on how the machine works.

Some performance-based assessments require students to work with others. Such tasks often include multimodal presentation of the stimuli and responses are not restricted to writing. While such performance-based assessments, including assessments

modeled after games, can create an environment where students acquire and demonstrate skills not typically captured in traditional assessments (e.g., communication and collaborative problem solving), the multimodal aspects of tasks and responses create challenges for measurement equivalence. Some newer assessment types also require students to interact with on-screen actors or avatars. Assessments with interactive elements may introduce behaviors that are not observed in traditional assessments (Zapata-Rivera & Bauer, 2012), such as the tendency to explore features of the assessment platform, or to engage in behaviors that push its boundaries (e.g., deliberately providing irrelevant responses), in addition to variations of sociocultural differences already observed in traditional assessments that may result in different response times, constructs being assessed, and measurement properties of the assessments.

Game-like elements (e.g., use of avatars in animations and actors in videos) that have been introduced into assessments to engage students present a different set of challenges. For example, a text item can refer to a “fellow student” without mentioning race, hair color, gender, age, or dressing style, whereas the use of avatars and actors will inevitably reflect these characteristics (Popp, Tuzinski, & Fetzer, 2016). Lee and Park (2011) found that minorities reported a lower sense of belonging and less desire to participate in a game with more white avatars than when the ethnic diversity of the avatars was more balanced. Such physical attributes can introduce social identity threats and affect student motivation (Baylor, 2011) and score comparability. The representation of diversity is often necessary and desired in state assessments with a diverse population and can be achieved in videos and animations.

Score comparability issues can also arise when sociocultural groups interpret pictures and videos differently based on their experiences. The extensive use of images in some new item types, such as hot-spot items (i.e., identification of correct or incorrect zones) and drag-and-drop image matching, requires that the images and videos represent the same notion to students from different sociocultural backgrounds (Solano-Flores & Nelson-Barber, 2001).

Despite the challenges, next-generation assessments can confer comparability advantages that are lacking in traditional assessments. Videos offer advantages in cases where some language and sociocultural groups might otherwise require test accommodations. In some cases, administering items through a video (or through audio) can be akin to the *human read-aloud* accommodation provided for some state assessments such as the LEAP 2025 Assessment, offered by the Louisiana Department of Education (Data Recognition Corporation, 2016), where items are read verbatim to individual students needing test accommodations. The readers may not clarify, provide additional information, assist, or influence the student’s response in any way. For mathematics read aloud, readers may read the title, provide a general overview of the image (e.g., graphs, equations), and describe the details in a succinct manner. The process requires the recruitment and the training of many readers. With video or audio test items, variation due to readers is eliminated as all examinees experience the same stimulus.

The current guidelines for test adaptations have yet to consider the possibilities and limitations of the new assessment types, partly due to the dearth of research studies. The decision to use a new format should be aligned to the assessment’s goals and fit with the intended construct and the ease of test adaptations. Also, in representing diversity, the designs of avatars should not default to stereotypes. A diverse group of

reviewers should evaluate the appropriateness of the diversity representation of avatars and actors. Finally, students' familiarity with technology should be considered (Ercikan, Asil, & Grover, 2018). Students who regularly use video chatting platforms or are accustomed to receiving news or information online may have an edge over students who have little or no access to computers outside the school environment.

## CONCLUSION

In this chapter, our goal has been to highlight the importance of following carefully designed procedures for designing and developing assessments for multilingual and multicultural contexts, establishing comparability of assessments and scores across language groups, and taking multiple societal factors into account in using and interpreting scores.

Comparability of scores for students from different language and sociocultural backgrounds is central to the validity of inferences from assessments. Validity is compromised when scores from multiple language versions of assessments are compared, implicitly or explicitly, without establishing comparability. For example, aggregating scores from different language versions of assessments, such as English and Spanish, at the class, school, or higher level makes an implicit assumption of score comparability of these language versions of assessments and the resulting scores. The incorrect assumption of score comparability compromises the validity of inferences when comparing students' performances at the class, school, or higher level. More explicit comparisons, such as comparing performance levels of students who took the assessment in English versus Spanish, may disguise or exaggerate performance differences when comparability has not been established.

Research on comparability issues in multiple language versions of assessments has been evolving to consider different assessment contexts. As assessments increasingly include multiple modes of administration (e.g., paper, digital, and multiple devices) and interactivity, assessment developers must also consider the expanded ways student backgrounds may contribute to how differently students engage with assessments and the resulting comparability issues across languages. It is important to highlight that the comparability of assessments and scores across languages is expected to be sensitive to population, cultural, and societal contexts. The validity and comparability evidence need to be updated periodically given potential changes in the society, education systems, and sociocultural context of assessments over the years.

## REFERENCES

- AERA (American Educational Research Association), APA (American Psychological Association), & NCME (National Council on Measurement in Education). (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.
- AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.
- Arffman, I. (2010). Equivalence of translations in international reading literacy studies. *Scandinavian Journal of Educational Research*, 54(1), 37–59.
- Arim, R., & Ercikan, K. (2005, April). *Comparability between the US and Turkish versions of the Third International Mathematics and Science Study's (TIMSS) mathematics test results*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.

- Bachman, H. J., Votruba-Drzal, E., El Nokali, N. E., & Castle Heatly, M. (2015). Opportunities for learning math in elementary school: Implications for SES disparities in procedural and conceptual math skills. *American Educational Research Journal*, 52(5), 894–923.
- Baylor, A. L. (2011). The design of motivational agents and avatars. *Educational Technology Research and Development*, 59(2), 291–300.
- Berliner, D. C. (2012). Effects of inequality and poverty vs. teachers and schooling on America's youth. *Teachers College Record*, 116(1). Retrieved from <http://www.tcrecord.org/Content.asp?ContentID=16889>.
- Blömeke, S., Suhl, U., Kaiser, G., & Döhrmann, M. (2012). Family background, entry selectivity and opportunities to learn: What matters in primary teacher education? An international comparison of fifteen countries. *Teaching and Teacher Education*, 28(1), 44–55.
- Brislin, R. W., Lonner, W. J., & Thorndike, R. M. (1973). *Cross-cultural research methods*. New York: Wiley.
- Byrne, B. M., & van de Vijver, F. J. R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing*, 10, 107–132.
- Campbell, L. (2003). How to show languages are related: Methods for distant genetic relationship. In B. D. Joseph & R. D. Janda (Eds.), *The handbook of historical linguistics* (pp. 262–282). Malden, MA: Blackwell Publishing.
- Carnevale, A. P., & Rose, S. J. (2003). *Socioeconomic status, race/ethnicity, and selective college admissions*. Century Foundation. Retrieved July 18, 2019, from <https://files.eric.ed.gov/fulltext/ED482419.pdf>.
- Cook, L. L. (2006). *Practical considerations in linking scores on adapted tests*. Keynote address at the 5th International Meeting of the International Test Commission, Brussels, Belgium.
- Cook, L. L., & Schmitt-Cascallar, A. P. (2005). Establishing score comparability for tests given in different languages. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 139–170). Mahwah, NJ: Lawrence Erlbaum Associates.
- Data Recognition Corporation. (2016). *LEAP 2025 accommodations and accessibility features user guide*. Retrieved May 30, 2019, from <https://www.louisianabelieves.com/docs/default-source/assessment/leap-accessibility-and-accommodations-manual.pdf?sfvrsn=12>.
- Derrida, J. (1998). *Monolingualism of the other, or, the prosthesis of origin*. Palo Alto, CA: Stanford University Press.
- Dorans, N. J., & Holland, P. W. (1992). DIF detection and description: Mantel-Haenszel and standardization. *ETS Research Report Series*, 1992(1), i–40.
- Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, 23(4), 355–368.
- Elder, T. E., Figlio, D. N., Imberman, S. A., & Persico, C. I. (2019). *School segregation and racial gaps in special education identification* (No. w25829). National Bureau of Economic Research.
- Ercikan, K. (1998). Translation effects in international assessments. *International Journal of Educational Research*, 29(6), 543–553.
- Ercikan, K. (2002). Disentangling sources of differential item functioning in multi-language assessments. *International Journal of Testing*, 2, 199–215.
- Ercikan, K. (2003). Are the English and French versions of the Third International Mathematics and Science Study administered in Canada comparable? Effects of adaptations. *International Journal of Educational Policy, Research and Practice*, 4, 55–76.
- Ercikan, K., Asil, M., & Grover, R. (2018). Digital divide: A critical context for digitally based assessments. *Education Policy Analysis Archives*, 26(51), 1–24. Retrieved from <https://epaa.asu.edu/ojs/article/view/3817>.
- Ercikan, K., Gierl, M. J., McCreith, T., Puhan, G., & Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education*, 17(3), 301–321.
- Ercikan, K., & Gonzalez, E. (2008, March). *Score scale comparability in international assessments*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.
- Ercikan, K., & Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing*, 5, 23–35.

- Ercikan, K., & Lyons-Thomas, J. (2013). Adapting tests for use in other languages and cultures. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology*, Vol. 3. *Testing and assessment in school psychology and education* (pp. 545–569). Washington, DC: American Psychological Association.
- Ercikan, K., & McCreith, T. (2002). Effects of adaptations on comparability of test items and test scores. In *Secondary analysis of the TIMSS data* (pp. 391–405). Dordrecht, the Netherlands: Springer.
- Ercikan, K., & Oliveri, M. E. (2013). Is fairness research doing justice? A modest proposal for an alternative validation approach in differential item functioning (DIF) investigations. In M. Chatterji (Ed.), *Validity, fairness and testing of individuals in high stakes decision-making context* (pp. 69–86). Bingley, UK: Emerald.
- Ercikan, K., Oliveri, M. E., & Sandilands, D. (2013). Large scale assessments of achievement in Canada. In J. A. C. Hattie and E. M. Anderman (Eds.), *The international handbook of student achievement* (pp. 456–459). New York: Routledge.
- Ercikan, K., & Roth, W.-M. (2006). What good is polarizing research into qualitative and quantitative? *Educational Researcher*, 35, 14–23.
- Ercikan, K., Roth, W.-M., & Asil, M. (2015). Cautions about uses of international assessments. *Teachers College Record*, 117(1), 1–28.
- Ercikan, K., Roth, W.-M., Simon, M., Sandilands, D., & Lyons-Thomas, J. (2014). Inconsistencies in DIF detection for sub-groups in heterogeneous language groups. *Applied Measurement in Education*, 27(4), 273–285.
- Florida Department of Education (n.d.). *Florida NAEP 2015 English language learners (ELL) inclusion guidelines*. Retrieved May 30, 2019, from <http://www.fldoe.org/core/fileparse.php/5423/urlt/ELLAGPPA.doc>.
- Gahungu, A., Gahungu, O., & Luseno, F. (2011). Educating culturally displaced students with truncated formal education (CDS-TFE): The case of refugee students and challenges for administrators, teachers, and counselors. *International Journal of Educational Leadership Preparation*, 6(2), 1–19. Retrieved from <https://eric.ed.gov/?id=EJ973832>.
- Gee, J. P. (2013). Reading as situated language: A sociocognitive perspective. In D. E. Alvermann, N. J. Unrau, & R. B. Ruddell (Eds.), *Theoretical models and processes of reading* (6th ed., pp. 136–151). Newark, DE: International Reading Association.
- Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment*, 6(4), 304.
- Gierl, M. J. (2000). Construct equivalence on translated achievement tests. *Canadian Journal of Education*, 25(4), 280–296.
- Gierl, M. J., Rogers, W. T., & Klinger, D. (1999, April). *Using statistical and judgmental reviews to identify and interpret translation DIF*. Presented at the annual meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.
- Greenfield, P. M. (1997). You can't take it with you: Why ability assessments don't cross cultures. *American Psychologist*, 52(10), 1115–1124.
- Grisay, A. (2003). Translation procedures in OECD/PISA 2000 international assessment. *Language Testing*, 20(2), 225–240.
- Grover, R. K., & Ercikan, K. (2017). For which boys and which girls are reading assessment items biased against? Detection of differential item functioning in heterogeneous gender populations. *Applied Measurement in Education*, 30(3), 178–195.
- Güzel, Ç. I., & Berberoglu, G. (2005). An analysis of the programme for international student assessment 2000 (PISA 2000) mathematical literacy data for Brazilian, Japanese and Norwegian students. *Studies in Educational Evaluation*, 31(4), 283–314.
- Haberman, S. J., & Dorans, N. J. (2011). Sources of score scale inconsistency. *ETS Research Report Series*, 2011(1), i–9.
- Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3–38). Mahwah, NJ: Lawrence Erlbaum Associates.
- Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (Eds.). (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Hambleton, R. K., & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Journal of Applied Testing Technology*, 1(1), 1–30.



- Holland, P. W., & Wainer, H. (2012). *Differential item functioning*. New York: Routledge.
- ITC (International Test Commission). (2018). ITC guidelines for translating and adapting tests (2nd ed.). *International Journal of Testing*, 18, 101–134.
- John-Steiner, V., & Mahn, H. (1996). Sociocultural approaches to learning and development: A Vygotskian framework. *Educational Psychologist*, 31(3–4), 191–206.
- Kirshner, B., Gaertner, M., & Pozzoboni, K. (2010). Tracing transitions: The effect of high school closure on displaced students. *Educational Evaluation and Policy Analysis*, 32(3), 407–429.
- Lee, J. E. R., & Park, S. G. (2011). "Whose second life is this?" How avatar-based racial cues shape ethnoracial minorities' perception of virtual worlds. *Cyberpsychology, Behavior, and Social Networking*, 14(11), 637–642.
- Lee, V. E., & Burkam, D. T. (2002). *Inequality at the starting gate: Social background differences in achievement as children begin school*. Washington, DC: Economic Policy Institute.
- Lipka, J., & McCarty, T. L. (1994). Changing the culture of schooling: Navajo and Yup'ik cases. *Anthropology & Education Quarterly*, 25(3), 266–284.
- López, A. A., Turkan, S., & Guzmán-Orth, D. (2017). Conceptualizing the use of translanguaging in initial content assessments for newly arrived emergent bilingual students. *ETS Research Report Series*, 2017(1), 1–12.
- Marotta, L., Tramonte, L., & Willms, J. D. (2015). Equivalence of testing instruments in Canada: Studying item bias in a cross-cultural assessment for preschoolers. *Canadian Journal of Education*, 38(3).
- McEniry, M., Samper-Ternent, R., Flórez, C. E., Pardo, R., & Cano-Gutierrez, C. (2019). Patterns of SES health disparities among older adults in three upper middle- and two high-income countries. *The Journals of Gerontology: Series B*, 74(6), 25–37.
- McQueen, J., & Mendelovits, J. (2003). PISA reading: Cultural equivalence in a cross-cultural study. *Language Testing*, 20(2), 208–224.
- Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, 18(2), 5–11.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741.
- New York State Education Department. (2016). *Testing accommodations for students with disabilities and English language learners*. Retrieved May 30, 2019, from <http://www.p12.nysed.gov/assessment/accommodations/testingaccomell-16.pdf>.
- Ohio Department of Education. (2018). *Accessibility features for students taking the paper-based Ohio's state tests*. Retrieved May 30, 2019, from <http://education.ohio.gov/getattachment/Topics/Testing/Accommodations-on-State-Assessments/OHAccessManual.pdf.aspx?lang=en-US>.
- Ohio Department of Education. (2019). *Revised assessment accommodations for English learners*. Retrieved May 30, 2019, from <http://education.ohio.gov/getattachment/Topics/Other-Resources/English-Learners/Revised-Assessment-Accommodations-for-English-Lear/Announcement-EL-Accommodations-and-OGT-Retakes.pdf.aspx?lang=en-US>.
- Oliveri, M. E., & Ercikan, K. (2011). Do different approaches to examining construct comparability in multilanguage assessments lead to similar conclusions? *Applied Measurement in Education*, 24(4), 349–366.
- Oliveri, M. E., Ercikan, K., & Zumbo, B. D. (2014). Effects of population heterogeneity on accuracy of DIF detection. *Applied Measurement in Education*, 27(4), 286–300.
- Oliveri, M. E., Olson, B. F., Ercikan, K., & Zumbo, B. D. (2012). Methodologies for investigating item- and test-level measurement equivalence in international large-scale assessments. *International Journal of Testing*, 12(3), 203–223.
- Oregon Department of Education. (2019). *2019-20 Oregon accessibility manual*. Retrieved September 4, 2019, from [https://www.oregon.gov/ode/educator-resources/assessment/Documents/accessibility\\_manual.pdf](https://www.oregon.gov/ode/educator-resources/assessment/Documents/accessibility_manual.pdf).
- Padilla, J.-L., Benítez, I., & Castillo, M. (2013). Obtaining validity evidence by cognitive interviewing to interpret psychometric results. *Methodology*, 9, 113–122.
- Perry, L., & McConney, A. (2010). Does the SES of the school matter? An examination of socioeconomic status and student achievement using PISA 2003. *Teachers College Record*, 112(4), 1137–1162.
- Popp, E. C., Tuzinski, K., & Fetzer, M. (2016). Actor or avatar? Considerations for selecting appropriate formats for assessment content. *Technology and Testing: Improving Educational and Psychological Measurement*, 79–103.



- Quealy, K., & Shapiro, E. (2019, June 17). Some students get extra time for New York's elite high school entrance exam. 42% are white. *The New York Times*. Retrieved from <https://www.nytimes.com/interactive/2019/06/17/upshot/nyc-schools-shsat-504.html>.
- Roth, W. M. (2009). Phenomenological and dialectical perspectives on the relation between the general and the particular. In K. Ercikan & W. M. Roth (Eds.), *Generalization in educational research* (pp. 235–260). New York: Routledge.
- Roth, W. M. (2010). *Language, learning, context: Talking the talk*. London, UK: Routledge.
- Roth, W. M., & Harama, H. (2000). (Standard) English as second language: Tribulations of self. *Journal of Curriculum Studies*, 32(6), 757–775.
- Shifrer, D., & Fish, R. (2019). A multilevel investigation into contextual reliability in the designation of cognitive health conditions among US children. *Society and Mental Health*. <https://doi.org/10.1177/2156869319847243>.
- Silva, S. M., Verhoeven, L., & van Leeuwe, J. (2011). Socio-cultural variation in reading comprehension development among fifth graders in Peru. *Reading and Writing*, 24, 951–969. <https://doi.org/10.1007/s11145-010-9242-2>.
- Sireci, S. G. (1997). Problems and issues in linking assessments across languages. *Educational Measurement: Issues and Practice*, 16(1), 12–19.
- Sireci, S. G. (2005). Using bilinguals to evaluate the comparability of different language versions of a test. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 117–138). Mahwah, NJ: Lawrence Erlbaum Associates.
- Sireci, S. G., & Allalouf, A. (2003). Appraising item equivalence across multiple languages and cultures. *Language Testing*, 20, 148–166.
- Sireci, S. G., Patsula, L., & Hambleton, R. K. (2005). Statistical methods for identifying flaws in the test adaptation process. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 93–115). Mahwah, NJ: Lawrence Erlbaum Associates.
- Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, 75, 417–453. <https://doi.org/10.3102/00346543075003417>.
- Smarter Balanced Assessment Consortium. (n.d.). *Embedded designated support—glossaries*. Retrieved August 9, 2019, from <https://portal.smarterbalanced.org/library/en/instructions-for-using-embedded-glossaries.pdf>.
- Snow, R. E. (1993). Construct validity and constructed-response tests. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 45–60). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Solano-Flores, G. (2006). Language, dialect, and register: Sociolinguistics and the estimation of measurement error in the testing of English language learners. *Teachers College Record*, 108(11), 2354.
- Solano-Flores, G., Backhoff, E., & Contreras-Niño, L. Á. (2009). Theory of test translation error. *International Journal of Testing*, 9(2), 78–91.
- Solano-Flores, G., Lara, J., Sexton, U., & Navarrete, C. (2001). *Testing English language learners: A sampler of student responses to science and mathematics test items*. Washington, DC: Council of Chief State School Officers.
- Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching*, 38(5), 553–573.
- Solano-Flores, G., Trumbull, E., & Nelson-Barber, S. (2002). Concurrent development of dual language assessments: An alternative to translating tests for linguistic minorities. *International Journal of Testing*, 2(2), 107–129.
- South Carolina Department of Education (n.d.). *SC READY online testing tools and supports*. Retrieved May 30, 2019, from [https://ed.sc.gov/scdoe/assets/File/districts-schools/special-ed-services/SC%20Ready%20Accommodations%20Charts\\_12-31-15.pdf](https://ed.sc.gov/scdoe/assets/File/districts-schools/special-ed-services/SC%20Ready%20Accommodations%20Charts_12-31-15.pdf).
- State of New Jersey Department of Education. (2019). *Testing accommodations*. Retrieved May 30, 2019, from <https://www.nj.gov/education/assessment/accommodations>.
- Tabaku, L., Carbuccia-Abbott, M., & Saavedra, E. (2018). *State assessments in languages other than English*. Retrieved August 8, 2019, from <https://files.eric.ed.gov/fulltext/ED590178.pdf>.
- Talento-Miller, E., Guo, F., & Han, K. T. (2013). Examining test speededness by native language. *International Journal of Testing*, 13(2), 89–104.

- Tanzer, N. K., & Sim, C. O. E. (1999). Adapting instruments for use in multiple languages and cultures: A review of the ITC guidelines for test adaptations. *European Journal of Psychological Assessment, 15*, 258–269.
- Tate, W. F. (1997). Race, ethnicity, SES, gender, and language proficiency trends in mathematics achievement: An update. *Journal for Research in Mathematics Education, 28*, 652–680.
- U.S. Census Bureau. (2017a). *2017 American Community Survey 1-year estimates*. Retrieved May 30, 2019, from [https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS\\_17\\_1YR\\_B16001&prodType=table](https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_17_1YR_B16001&prodType=table).
- U.S. Census Bureau. (2017b). *Characteristics of people by language spoken at home*. Retrieved May 30, 2019, from [https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS\\_14\\_5YR\\_S1603&prodType=table](https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_14_5YR_S1603&prodType=table).
- Ward, M. E., Shelley, K., Kaase, K., & Pane, J. F. (2008). Hurricane Katrina: A longitudinal study of the achievement and behavior of displaced students. *Journal of Education for Students Placed at Risk, 13*(2–3), 297–317.
- Wu, A., & Ercikan, K. (2006). Using multiple-variable matching to identify cultural sources of differential item functioning. *International Journal of Testing, 6*, 287–300.
- Zapata-Rivera, D., & Bauer, M. (2012). Exploring the role of games in educational assessment. In M. C. Mayrath, J. Clarke-Midura, D. H. Robinson, & G. Schraw (Eds.), *Technology-based assessments for twenty-first-century skills: Theoretical and practical implications from modern research* (pp. 147–169). Charlotte, NC: Information Age Publishing.
- Zhang, Y., Dorans, N. J., & Matthews-López, J. L. (2005). Using DIF dissection method to assess effects of item deletion. *ETS Research Report Series, 2005*(2), i–11.
- Zirkel, P. A., & Weathers, J. M. (2016). K–12 students eligible solely under section 504: Updated national incidence data. *Journal of Disability Policy Studies, 27*(2), 67–75.
- Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language Testing, 20*(2), 136–147.



# Interpreting Test-Score Comparisons

Randy E. Bennett, *Educational Testing Service*

## CONTENTS

INTRODUCTION .....	227
BASIC PREMISES .....	227
WEAKER COMPARISONS .....	228
Instruments That Are Nominally the Same .....	228
Different Instruments .....	229
Different Populations .....	230
SUGGESTIONS FOR PRACTICE .....	230
CONCLUSION .....	234
REFERENCES .....	234

## INTRODUCTION

This chapter focuses on making sense from test-score comparisons. The chapter begins with some basic premises. It then proceeds to a discussion of factors that can weaken the tenability of test-score comparisons. Finally, the chapter offers some suggestions for responsibly interpreting and communicating comparisons. Much of the content draws upon ideas and examples from preceding chapters.

## BASIC PREMISES

This chapter proceeds from the premise that getting meaning from assessment results inevitably requires some type of comparison. Without a benchmark or reference point, an assessment result can become an uninterpretable abstraction. To lend meaning to the results for an individual, the results may be referenced, or compared, to those of other test takers, to past performance, to the types of tasks that characterize performance at a particular score level, or to some absolute standard like a cut point indicative of broader domain proficiency.

Not only does deriving meaning from assessment results require some type of comparison, but some common comparative frame is usually needed for results to be

aggregated. That is, we cannot sensibly compute an average score for a group unless each member of that group has a result that is *comparable*.<sup>1</sup>

Comparisons are strongest when the same measure is given under substantively the same conditions to analogous student samples at the same point in time. In the case of comparisons of performance to an absolute standard, the similarities of conditions, student sample, and time point are with the conditions, time point, and student group assumed in setting the cut point. Comparisons become weaker as the measure, assessment conditions, student samples, or the time of administration begin to diverge. The more severe and numerous the divergences, the less defensible the comparison is likely to be.<sup>2</sup>

As defined above, strong comparisons will necessarily be limited to a subset of the comparisons assessment users may want to make. For that reason, it is important to identify each source of divergence and how that divergence might affect the tenability of the comparison.

### WEAKER COMPARISONS

In this section, three types of divergence are briefly discussed. They are divergence due to instruments (i.e., assessments) that are nominally the same, to dissimilar instruments, and to different examinee populations.

#### Instruments That Are Nominally the Same

The “same” instrument can, in practice, appear in several different forms. Each of those forms can introduce divergences that weaken our ability to make comparisons. There are at least three senses in which an instrument can appear in nominally different forms. One sense is literal and refers to the presentation of examination content. For example, the Programme for International Student Assessment (PISA) delivered its 2015 science examination in 90 different language versions (OECD, 2018). Comparisons across language versions pose challenges because ideas are not always directly translatable in forms that are similar in meaning, vocabulary level, or syntactic complexity, potentially affecting the difficulty of questions (see Chapter 8, Comparability in Multilingual and Multicultural Assessment Contexts). Moreover, the same content may require relatively little text to represent it in one language but a lengthier exposition in another language, differentially affecting reading demand.

The literal form of an assessment can change via the method chosen for its delivery: paper or computer. That change may be relatively minor, as when the multiple-choice questions from a paper test are presented in similar fashion on screen. The change is more significant, however, if the online version employs item types that the paper test does not (e.g., technology-enhanced items or simulation tasks) or if the modes of

---

<sup>1</sup> Group-score assessments that use direct estimation are an exception (e.g., National Assessment of Educational Progress [NAEP] and Programme for International Student Assessment [PISA]). Such estimation is, however, not used when individuals must be awarded scores, such as on state assessments.

<sup>2</sup> Somewhat different considerations apply when the same student is tested repeatedly over time to measure growth, for example, through annual state assessments placed on a vertical scale. These considerations might include the types of scores compared, the effectiveness of the scaling, and the degree to which the tested constructs overlap.

response are substantially different (e.g., answering an essay question on paper versus on a computer). These more significant differences may affect the difficulty of questions and perhaps even the skills measured (Bennett, 2003; Bennett et al., 2008; Horkay, Bennett, Allen, Kaplan, & Yan, 2006).

The same instrument can also take a different literal form through the provision of accommodations for students with disabilities or for English learner students (see Chapter 6, *Comparability When Assessing English Learner Students*, and Chapter 7, *Comparability When Assessing Individuals with Disabilities*). An obvious example would be the translation of the examination into Braille or provision of portions of the test in American Sign Language.

An examination also can take a different form when the presentation of the assessment is literally the same but there is divergence in how constructed-response questions are scored. A good example is found in the Smarter Balanced Assessment Consortium, in which member states choose a scoring vendor and whether that vendor uses human grading, machine grading, or both methods. Those choices may not necessarily produce the same results across states even within the same method, depending on the resolution procedures used when raters disagree or upon the particular machine-scoring algorithm that is employed (Bennett & Zhang, 2016).

The third way an instrument can take a different form is less obvious. This type of divergence occurs when the instrument's presentation, response mode, and scoring are, from an objective perspective, the same for all examinees. However, for any pair of examinees that have contrasting characteristics, that assessment may appear to be as different as night and day. Consider the following pairs: (1) an examinee with sight and one with visual impairment, each presented with an unaccommodated test; (2) an examinee who routinely composes essays on a computer and one who typically writes on paper, each given an online writing examination; (3) a native English speaker and an English learner, both taking the test in English; (4) two individuals, one from the mainstream culture and the second from an environment having very different practices, both presented with a reading comprehension test presuming significant background knowledge of U.S. cultural norms; (5) two examinees who are otherwise the same, but one has seen and practiced the test items in advance; (6) two comparable examinees with the exception that only one perceives the test's consequences to be personally significant; and (7) one student having received instruction and the other having not had sufficient opportunity to learn. In all these cases, how the examinees perform and the scores they receive are facts. However, the interpretations we give could be very different and, thereby, the comparisons between those examinees (and the groups to which they belong) are weakened.

### **Different Instruments**

In addition to divergence related to instruments that are nominally the same, comparisons can become weaker when performance on two different instruments is involved. The source of the weakened comparison is that different instruments will typically diverge in terms of the content and processes they measure, as well as the reference frames used to characterize performance.

This type of divergence occurs with some frequency. It can occur for assessment systems being used for the same purpose, such as when we try to compare the percentages



of students achieving proficiency on Smarter Balanced with those on the Partnership for Assessment of Readiness for College and Careers (PARCC) assessments. Such divergence also occurs between two different stand-alone tests used for the same purpose. One example would be use of the SAT or the ACT, and the TOEFL iBT® or the International English Language Testing System™, in making postsecondary admissions decisions; another example encompasses the many assessments used by states for classifying students as English learners (see Chapter 6, *Comparability When Assessing English Learner Students*). Finally, divergence can occur when measures built for one purpose are also used for and compared to assessments built for another purpose. Comparing the percentage of high school students who achieve proficiency when taking the ACT or the SAT as a state accountability measure to the analogous percentage taking an assessment built to measure a state's content standards directly might be an example (NCME, 2019).

### **Different Populations**

Finally, comparisons become weaker when the same instrument is administered to two student samples that diverge enough from one another that they can be considered as coming from different populations (where the intent is not to compare those different populations). An infamous example is the U.S. Department of Education's attempt to evaluate school achievement across states by using ACT and SAT performance (Wainer, Holland, Swinton, & Wang, 1985). That comparison was undermined by the fact that considerably different proportions of high school students took those tests in each state. A second example is when student performance is compared across states that have different accommodation policies for students with disabilities or English learners. A last example is when the same test is administered at two points in time and the population's composition has materially changed over that period (see Chapter 3, *Comparability of Aggregated Group Scores on the "Same Test"*).

## **SUGGESTIONS FOR PRACTICE**

In this section, we offer suggestions for interpreting score comparisons, discussing each one in turn.

A first step is to determine why a comparison might need to be made. The wisdom of making a comparison may vary with decision-making purpose so it is important to be clear about that purpose. Comparisons can be purely descriptive, made simply for reporting what occurred. An example is in detailing how various states are ranked in terms of their students' performance on the National Assessment of Education Progress (NAEP)—a matter of fact. In practice, this is the view held by some test sponsors, who choose to report results without interpretation. For example, NAEP reports typically stay quite close to the observed results.

Descriptive purposes can, however, quickly turn (or be turned) into inferential ones because we naturally want, and often automatically do, imbue facts with interpretation. Those interpretations, by definition, entail inferences, which together provide the basis for using results in decision making.

Interpretation is, in fact, what state policy makers, the press, and the public do with the descriptive results that come from NAEP. One or more of those groups could, for

example, infer that the observed differences among states were due to differences in teacher competency, the rigor of state education standards, the policies and practices for teacher evaluation, the population demographics, or some combination of factors. Each of these inferences, of course, has particular action implications. However, a more reasoned approach is to regard descriptive results as an opportunity for posing questions that, in turn, motivate the generation of additional evidence to help distinguish among competing interpretations.

A second basic step for interpreting test-score comparisons is to ascertain what methods might have been used to make the desired comparisons tenable. Comparisons can often be made more defensible by using statistical techniques as part of generating assessment results (e.g., making adjustments to allow scores from one test form to be compared with those from another test form). Methods for facilitating comparability vary in their requirements and the degree to which they produce exchangeable scores. As a consequence, some methods may be more suitable for particular decision-making purposes than others. Equating, concordance, and prediction are examples that range from stronger to weaker in their requirements and in the results that they produce. Other chapters in this volume describe these methods (see Chapter 2, Comparability of Individual Students' Scores on the "Same Test," and Chapter 5, Comparability Across Different Assessment Systems), as well as related technical concerns (see Chapter 4, Comparability Within a Single Assessment System). For interpreting score comparisons, we suggest identifying whether the method used (if any) supports the desired comparison.

A third step is to consider how and to whom results will be reported and how comparative claims will be made. Comparative claim statements can appear (or be implied in) score reports, press releases, websites, and other communications, all of which afford opportunities to help audiences make sensible comparisons and avoid untenable ones.

In preparing to report comparative results, it is best to determine first whether the same test was used, and whether it was administered under the same conditions to comparable student samples at the same point in time. If these circumstances do not hold, the specific divergence(s) should be identified and the impact of those divergences on the meaning of assessment results evaluated to the extent feasible. Many methods exist for evaluating the invariance of score meaning across different test variations (e.g., languages or delivery media), examinee populations, and administrative conditions (see Chapter 7, Comparability When Assessing Individuals with Disabilities, and Chapter 8, Comparability in Multilingual and Multicultural Assessment Contexts). A justification for making the comparison in the presence of those divergences should be offered, including a logical rationale and a delineation of the empirical evidence supporting or challenging the comparison.

Technical advisory committee (TAC) guidance is essential in considering the comparison, empirically evaluating its tenability, and creating a justification built on logic and evidence. Of central importance is to start from the premise that results have to be reported and that score comparisons will inevitably be made. The task then becomes one of fashioning communications that responsibly describe results, offer defensible comparisons, and warn against unwarranted inferences.

To that end, we suggest working with the TAC to adjust the strength of the comparative claim as a function of (1) the extent to which the instruments, assessment conditions, student samples, and time between administrations diverge, and (2) the extent of the logical and empirical support available to back the claim. Claim statements can

be adjusted in terms of confidence level based on these two factors. A high-confidence claim would be one for which there is no or little divergence, or there is some divergence but good justification for the comparison given that divergence (e.g., scores have been equated). A lower-confidence claim might be very plausible given current education theory but have limited or no empirical backing. Claims of this type should be more tentatively stated. In all cases, the caveats that attend to the comparison should be clearly articulated and unjustified inferences identified as such (Toulmin, 1958).

Table 9-1 gives some examples of possible comparisons along with more and less defensible claims related to them. Note that the more defensible claims stick closely to the measures used and populations assessed; tenability decreases as claims take on greater levels of generality. For example, it would be reasonable to claim that females scored higher than males on the 2011 NAEP eighth grade writing assessment when composing online essays on demand to persuade, explain, or convey experience. It would also be reasonable to suggest that U.S. eighth grade females were better writers than males *in that context*. Less tenable would be the claim that females were better writers than males generally because, among other things, these 2011 NAEP results targeted a single grade, composition in a particular medium (on computer), writing on demand (which may differ from classroom composition), and particular writing purposes. More general still, and quite untenable, would be the claim that females received better writing instruction than males, a causal attribution that NAEP is not designed to support (NCES, n.d.).

**TABLE 9-1** Example Comparisons and Claims of Varying Degrees of Defensibility

Comparison	Well-Supported Claim	Claim Requiring Additional Evidence	Claim Not Recommended
Performance of male and female students on eighth grade 2011 NAEP writing assessment	Female students scored higher than male students at the eighth grade level when composing online essays on demand to persuade, explain, or convey experience	Female students write better than male students <i>Comment:</i> This comparison requires evidence that the NAEP results extend to other grades, to writing on paper, and to other writing purposes than those assessed	Female students received better writing instruction than male students <i>Comment:</i> This comparison presumes a causal connection between the instruction received and the outcome measured, which NAEP was not designed to support

TABLE 9-1 Continued

Comparison	Well-Supported Claim	Claim Requiring Additional Evidence	Claim Not Recommended
Performance of students in the same school taking the fourth grade state reading assessment in 2018 and 2019	The percentage of fourth grade students reaching proficiency increased by 10 points from 2018 to 2019	Fourth grade reading instruction is having a positive effect <i>Comment:</i> This claim would be strengthened by evidence that the two assessed fourth grade populations were demographically comparable, similar percentages of eligible students tested, the test did not change in any material way across the 2 years, no pre-knowledge or other forms of cheating were evident, and no errors in scoring or analysis occurred	The reading skills of fourth graders improved <i>Comment:</i> We do not know that the fourth graders improved because the same group of students was not compared. This claim might be better stated as, "The reading skills of the 2019 fourth grade students were greater than those of the 2018 fourth graders"
Performance of two third grade students, each taking their home district's interim assessment	Both students received the same percentile score in mathematics and are estimated to be equally competent with respect to other third graders in the respective tests' norming samples	The students have similar levels of mathematics competency <i>Comment:</i> This claim would be strengthened by evidence that the two assessments were built to the same content standards, had similar types of items covering those standards to comparable degrees and levels of rigor, used similar student samples and methods in setting scales and norms, and were administered under analogous conditions	Their districts are equally effective in educating them <i>Comment:</i> This claim presumes that the districts' efforts are solely responsible for the students' achievement, goes well beyond mathematics, assumes that the districts have offered equal opportunities to learn, and is based on a single achievement indicator
Performance of 10th grade students on a new state achievement test compared to last year's results on the old test	This year's cohort had a lower percentage proficient	This year's test is harder <i>Comment:</i> This claim would be strengthened by evidence that the proficiency standards for the two tests were set in ways that allow meaningful comparison and there were no material changes in the 10th grade populations	The state's students are becoming less intellectually capable <i>Comment:</i> This claim conflates achievement of content standards with intellectual capability and presumes that differences between the two measurements are rooted in the populations measured rather than changes in the test

continued

TABLE 9-1 Continued

Comparison	Well-Supported Claim	Claim Requiring Additional Evidence	Claim Not Recommended
Performance of a school's fourth grade students on its English language arts (ELA) state test to an estimate of the national average for all fourth grade students taking their respective state ELA tests, when those tests are rescaled through NAEP	The school's fourth graders scored below the national average in ELA	The ELA achievement of the school's fourth graders is below the national average <i>Comment:</i> This claim would be strengthened by evidence that ELA content standards were similar enough across states to allow for creating a coherent common scale, the scaling was technically adequate, and assessment participation rates and accommodation policies were not divergent from the national average	Educational opportunity in the school is below the national average <i>Comment:</i> This claim presumes that an outcome, test performance, is equivalent to an input, opportunity

## CONCLUSION

This chapter focused on interpreting test-score comparisons. The chapter began with the premise that comparisons are inevitable and, in fact, desirable because obtaining meaning from assessment results requires them. We noted that comparisons are strongest when the same measure is given under substantively the same conditions to comparable student samples at the same point in time, with departures serving to weaken comparisons. The more severe and numerous the departures, the less defensible the comparison is likely to be. In interpreting test-score comparisons one should articulate why the comparison is being made, ascertain if the comparison is appropriate given the technical methods used, present a rationale based on logic and evidence to support the comparison, and warn audiences against inappropriate inferences. Finally, tenable comparisons will usually be ones that stay reasonably close to the measures employed and populations tested. As comparative claims become more general, their reasonableness usually declines.

## REFERENCES

- Bennett, R. E. (2003). *Online assessment and the comparability of score meaning* (Research Memorandum 03-05). Princeton, NJ: ETS.
- Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & Yan, F. (2008). Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *Journal of Technology, Learning and Assessment*, 6(9). Retrieved from <http://files.eric.ed.gov/fulltext/EJ838621.pdf>.
- Bennett, R. E., & Zhang, M. (2016). Validity and automated scoring. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 142–173). New York: Routledge.
- Horkay, N., Bennett, R. E., Allen, N., Kaplan, B., & Yan, F. (2006). Does it matter if I take my writing test on computer? An empirical study of mode effects in NAEP. *Journal of Technology, Learning and Assessment*, 5(2). Retrieved from <http://files.eric.ed.gov/fulltext/EJ843858.pdf>.

- NCES (National Center for Education Statistics). (n.d.). *Interpreting NAEP technology and engineering literacy results*. Retrieved from [https://nces.ed.gov/nationsreportcard/tel/interpret\\_results.aspx#cautions](https://nces.ed.gov/nationsreportcard/tel/interpret_results.aspx#cautions).
- NCME (National Council on Measurement in Education). (2019). *National Council on Measurement in Education position statement on the use of college admissions test scores as academic indicators in state accountability systems*. Retrieved from [https://higherlogicdownload.s3.amazonaws.com/NCME/c53581e4-9882-4137-987b-4475f6cb502a/UploadedImages/Documents/Admission\\_Statement\\_06-16-19.pdf](https://higherlogicdownload.s3.amazonaws.com/NCME/c53581e4-9882-4137-987b-4475f6cb502a/UploadedImages/Documents/Admission_Statement_06-16-19.pdf).
- OECD (Organisation for Economic Co-operation and Development). (2018). *PISA 2015 science test*. Retrieved from <http://www.oecd.org/pisa/test/other-languages>.
- Toulmin, S. (1958). *The uses of argument*. Cambridge, UK: Cambridge University Press.
- Wainer, H., Holland, P. W., Swinton, S., & Wang, M. H. (1985). On "State education statistics." *Journal of Educational Statistics*, 10, 293–325.





## Biographical Sketches of Steering Committee Members and Authors

**Randy E. Bennett** is the Norman O. Frederiksen Chair in Assessment Innovation in the Research & Development Division at Educational Testing Service in Princeton, New Jersey. Bennett's work has focused on integrating advances in cognitive science, technology, and educational measurement to create approaches to assessment that have positive impacts on teaching and learning. From 1999 through 2005, he directed the National Assessment of Educational Progress Technology Based Assessment project, which included the first administration of computer-based performance assessments with nationally representative samples of school students and the first use of "clickstream," or logfile, data in such samples to measure the processes used in problem solving. From 2007 to 2016, he directed an integrated research initiative titled Cognitively-Based Assessment of, for, and as Learning (CBAL), which focused on creating theory-based summative and formative assessment intended to model good teaching and learning practice. Bennett is the immediate past president of the International Association for Educational Assessment (2016–2019), an organization primarily constituted of governmental and nongovernmental nonprofit measurement organizations throughout the world, and the past president of the National Council on Measurement in Education (NCME) (2017–2018), whose members are individuals employed primarily in universities, testing organizations, state departments of education, and school districts. He is a fellow of the American Educational Research Association, winner of the NCME Bradley Hanson Contributions to Educational Measurement Award, and winner of the Distinguished Alumni Award from Teachers College, Columbia University.

**Amy I. Berman** is the deputy director of the National Academy of Education, where she works to advance the strategic and research initiatives for the more than 200 member organizations. She also is an adjunct professor at The George Washington University Graduate School of Education and Human Development (GSEHD). Prior to that she was an education civil rights lawyer as an enforcement director at the U.S. Department

of Education's Office for Civil Rights and the section chief at the U.S. Department of Justice, Civil Rights Division, Educational Opportunities Section. In those positions, she worked to ensure equal access to education through the vigorous enforcement of civil rights laws, including in the areas of race, national origin, sex, religion, disability, and language. In addition to enforcement, she worked on key guidance documents addressing the use of race in schools, harassment in schools, education of English learner students, and the requirement to educate all students, regardless of immigration status. She also held a fellowship addressing prisoners' civil rights issues, worked as an associate at Sullivan & Cromwell, and clerked for the late Honorable Robert J. Ward of the U.S. District Court for the Southern District of New York. She has been an adjunct professor at The George Washington University School of Law and the American University Washington College of Law. She is a doctoral candidate in Education Policy at GSEHD, and she holds a J.D. from Harvard Law School and a B.S. from Cornell University. Recently, she co-edited a volume of *The ANNALS of the American Academy of Political and Social Science*, "What Use Is Educational Assessment?"

**Charles A. DePascale** is an educational consultant specializing in technical and policy issues related to the design of K–12 assessment systems and the interpretation and use of educational assessments by educators and policy makers. As a senior associate at the National Center for the Improvement of Educational Assessment from 2002 to 2019, he provided technical guidance and support in the design and use of assessment systems to support student, educator, and school accountability systems through direct work with individual states as well as through participation in multi-state assessment programs and research projects, technical advisory committees, conference presentations, workshops, blog posts, and occasional papers. Over the past decade he has worked extensively on issues related to the comparability of assessment results under various conditions, including comparability across alternate forms within a state assessment program, comparability across tests administered via different devices, and comparability across states within a multi-state assessment program.

**Kadriye Ercikan** is the vice president of psychometrics, statistics, and data sciences at the Educational Testing Service (ETS) and a professor emerita at the University of British Columbia. She is responsible for data analysis and psychometric support of ETS's major testing products and contracts, as well as for foundational and applied statistical and psychometric research. Her research focuses on designing and validating assessments of complex thinking, the assessment of linguistic minorities, and fairness and validity issues in cross-cultural and international assessments. Ercikan is a fellow of the International Academy of Education. Her research has resulted in 6 books, 1 special issues of refereed journals, and more than 100 publications. One co-edited book, *Validating Score Meaning in the Next Generation of Assessments*, was selected for publication as part of the National Council on Measurement in Education book series. She was also awarded the American Educational Research Association Division D Significant Contributions to Educational Measurement and Research Methodology recognition for another co-edited volume, *Generalizing from Educational Research: Beyond Qualitative and Quantitative Polarization*, and received an Early Career Award from the University of British Columbia.

**Molly Faulkner-Bond** is a senior research associate at WestEd, where her work focuses on English learners (ELs), policy, and assessment. In this role, Faulkner-Bond provides technical assistance and advising to a number of state education agencies (SEAs) around issues such as evidence-based practices for EL instruction and program design; valid use and interpretation of scores from large-scale standardized assessments for different purposes; and systems and processes for monitoring and evaluating policies and programs. Across all of her technical assistance work, she focuses on distilling the research base into clear, actionable information that will help SEA administrators to grow their own understanding and make more informed decisions to support their teachers and students. Prior to joining WestEd, Faulkner-Bond was a grant program officer at the Institute of Education Sciences, where she provided technical assistance and monitoring to applicants and recipients of multi-year research grants focused on improving educational opportunities and outcomes for ELs. She worked previously at the Educational Testing Service in the English language learning and assessment division, and began her career at a federal contracting firm that provided technical assistance around the collection of validity evidence related to ELs and students with disabilities for federal peer review. She has co-authored a book on federal policies affecting ELs, co-edited a book on educational measurement and assessment, and co-authored several articles on assessment validity and score reporting for both ELs and the general population.

**Louis Gomez** is a social scientist dedicated to educational improvement. His research and design efforts are aimed at helping to support community formation in schools and other organizations so that they can collaboratively create new approaches to teaching, learning, and assessment. With colleagues, he has worked to bring networked-based improvement science to the field of education. This work is aimed at helping the field take a new perspective on design, educational engineering, and development efforts that catalyze long-term, cooperative initiatives. The work gains much of its power because it is carried out in highly focused collaboratives that Gomez and colleagues call networked improvement communities. Gomez is a professor of education at the University of California, Los Angeles, and a senior fellow at the Carnegie Foundation for the Advancement of Teaching.

**Brian Gong** has been involved with comparability as an aspect of assessment validity and utility for more than 30 years. His focus on validation has been broadly reflected in his work, including the nature of claims, design of test blueprints that will support sufficient evidence to support claims, design of performance and portfolio assessments, means to support accurate scoring of complex student responses, and guides for design and evaluation of assessment systems, such as alignment methodologies. He has worked with several applications where achieving comparability requires innovative designs and evaluation procedures, including alternate assessments for students with severe cognitive disabilities, complex performance assessments, computer adaptive assessments, assessments with local administration and scoring, assessments of non-cognitive skills and dispositions, and assessments of complex domains such as the Next Generation Science Standards. He has extensive experience in the challenges to achieving comparability in large-scale testing, having worked extensively supporting states to implement their assessment and accountability systems in practical ways, from

inclusion to linking to transitioning vendors and tests to reporting through producing documentation that will be approved through technical advisory committees and federal peer review. Gong is a senior associate at the National Center for the Improvement of Educational Assessment, a nonprofit consulting organization based in Dover, New Hampshire, that provides technical support to states and other organizations around assessment and accountability. Prior to co-founding the Center, Gong served as the associate commissioner for curriculum, instruction, and assessment in the Kentucky Department of Education and as a senior research scientist at the Educational Testing Service.

**Edward Haertel** is the Jacks Family Professor of Education, emeritus, at Stanford University, where his research and teaching focused on quantitative research methods, psychometrics, and educational policy, especially test-based accountability and the use of test data for educational program evaluation. Haertel's early work investigated the use of latent class models for item response data. His more recent projects have included studies of standard setting and standards-based score interpretations, statistical properties of test-based accountability systems, metric-free measures of score gaps and trends, and the examination of value-added models for teacher evaluation from a psychometric perspective. Representative publications include *Reliability and Validity of Inferences About Teachers Based on Student Test Scores* (14th William H. Angoff Memorial Lecture, 2013); *Selection of Common Items as an Unrecognized Source of Variability in Test Equating* (2014, with M. Michaelides); *Fairness Using Derived Scores* (2016, with A. Ho); *Engaging Methodological Pluralism* (2016, with P. A. Moss); *Tests, Test Scores, and Constructs* (2018); and *Measuring Cultural Dimensions of Classroom Interactions* (2018, with B. Jensen and S. Grajeda). Haertel has served as president of the National Council on Measurement in Education, chaired the Technical Advisory Committee concerned with the design and evolution of California's test-based school accountability system, is a past chair of the National Academies of Sciences, Engineering, and Medicine's Board on Testing and Assessment, and is a former member of the National Assessment Governing Board.

**Larry V. Hedges** is the chairman of the Department of Statistics, the Board of Trustees Professor of Statistics, Psychology, in the School of Education and Social Policy, and in the Institute for Policy Research at Northwestern University. He is best known for his work on statistical methods for meta-analysis and its applications to evidence-based policy. He has also worked on the design of social experiments, the assessment of student achievement nationally and cross-nationally, and the role of uncertainty in basic models for cognition in psychology. He has authored or co-authored numerous journal articles and 11 books. He is an elected member of the National Academy of Education and is a fellow of the American Academy of Arts & Sciences, the American Statistical Association, the American Psychological Association, and the American Educational Research Association. He also served as chair of the Board of Directors of the National Board for Education Sciences. Hedges has served as the quantitative methods editor of *Psychological Bulletin*, the founding co-editor of the *Journal of Research on Educational Effectiveness*, and the editor of the *Journal of Educational and Behavioral Statistics*.

**Joan Herman** is the director, emerita, of the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) at the University of California, Los Angeles

(UCLA). Her research has explored the effects of testing on schools and the design of assessment systems to support school planning and instructional improvement. Her recent work focuses on teachers' formative assessment practices and fairness in classroom and large-scale assessment. She also has wide experience as an evaluator of school reform. Herman is noted for bridging research and practice. Among her books are the *Turnaround Toolkit* and *A Practical Guide to Alternative Assessment*, both of which have been popular resources for schools across the country. A former teacher and school board member, Herman also has published extensively for research, practitioner, and policy audiences on evaluation and assessment topics. She is the past president of the California Educational Research Association; has held a variety of leadership positions in the American Educational Research Association (AERA), has provided leadership for Standards for Educational and Psychological Testing, especially in the area of fairness, and serves as the editor of *Educational Assessment*. Nationally recognized as a leader in the field, Herman has been honored as an elected member of the National Academy of Education, as a fellow of the AERA, and received numerous awards for excellence. Herman received her B.A. in sociology from the University of California, Berkeley, was awarded an M.A. and an Ed.D. in learning and instruction from UCLA and is a member of Phi Beta Kappa.

**Leslie Keng** is a senior associate at the National Center for the Improvement of Educational Assessment. He has more than a decade of experience supporting states and consortia in the development, implementation, and evaluation of assessment and accountability systems through empirical and evidence-based approaches. He has helped several states with their transitions in their assessment and accountability systems by offering guidance and technical consultation through significant changes. These changes included moving from a consortium-based assessment to a custom state-developed solution, transitioning to new assessment vendors, and implementing new school accountability models based on legislative requirements. Prior to joining the Center, Keng was a principal research scientist at Pearson. During his 11 years at Pearson, he supported two of the largest testing programs in the United States—in Texas (STAAR EOC) and the Partnership for Assessment of Readiness for College and Careers (PARCC) as lead psychometrician. He helped launch the next-generation assessment systems for both programs by overseeing psychometric tasks and providing technical support during all phases of the testing development process. Keng is also one of the architects of the evidence-based standard setting method, used to set performance standards in a number of assessment programs, including in New York, Texas, and PARCC. A former high school mathematics teacher, Keng earned a bachelor's degree in computer science from the University of Waterloo and a bachelor of education from Queen's University in Canada. He also completed a master's degree in statistics and received his Ph.D. in educational psychology (quantitative methods) from The University of Texas in Austin.

**Scott Marion** is the president and executive director of the National Center for the Improvement of Educational Assessment. He is a national leader in designing innovative and comprehensive assessment systems to support instructional and accountability uses and is working to better conceptualize and implement high-quality, balanced



systems of assessment and accountability. He is actively engaged with a broad range of Center clients, including chief state school officers, legislators, state and district assessment and accountability leaders, and classroom teachers. His projects include designing and supporting states in implementing assessment and accountability initiatives, providing technically defensible policy guidance, and implementing high-quality, locally designed, performance-based assessments. Marion coordinates and/or serves on six state or district Technical Advisory Committees for assessment, accountability, and educator evaluation. He has served on multiple National Academies of Sciences, Engineering, and Medicine committees, including to support designs for the next-generation science assessments, investigating the issues and challenges associated with incorporating value-added measures in educational accountability systems, and outlining best practices in state assessment systems. Marion has published dozens of articles and chapters in peer-reviewed journals and edited volumes; he also regularly presents his work at national conferences. Prior to joining the Center in early 2003, Marion was the director of assessment and accountability for the Wyoming Department of Education. He also serves his community as a member of the Rye (NH) School Board. Marion received a Ph.D. from the University of Colorado Boulder with a concentration in measurement and evaluation.

**Maura O’Riordan** is a doctoral student in the Research in Educational Measurement and Psychometrics program within the College of Education at the University of Massachusetts. She earned her master’s degree in education from Simmons College and her bachelor’s degree in mathematics and education from St. Michael’s College. She was a special education high school math teacher prior to beginning her doctoral studies. Her research focus is in testing accommodation use and the assignment of accommodations, as well as test development procedures.

**James W. Pellegrino** is the Liberal Arts and Sciences Distinguished Professor and the co-director of the Learning Sciences Research Institute at the University of Illinois at Chicago. His research and development interests focus on children’s and adult’s thinking and learning and the implications of cognitive research and theory for assessment and instructional practice. He has published more than 300 books, chapters, and articles in the areas of cognition, instruction, and assessment. His research is funded by the National Science Foundation, the Institute of Education Sciences, and private foundations. He has served on several National Academies of Sciences, Engineering, and Medicine study committees, including chair of the Committee for the Evaluation of the National and State Assessments of Educational Progress, co-chair of the Committee on Learning Research and Educational Practice, and co-chair of the Committee on the Foundations of Assessment, which issued the report *Knowing What Students Know: The Science and Design of Educational Assessment*. Most recently he served as a member of the Committee on Science Learning: Games, Simulations and Education, as a member of the Committee on a Conceptual Framework for New Science Education Standards, as chair of the Committee on Defining Deeper Learning and 21st Century Skills, and co-chair of the Committee on Developing Assessments of Science Proficiency in K–12. He is a past member of the National Academies’ Board on Testing and Assessment, a

lifetime associate of the National Academy of Sciences, and a lifetime member of the National Academy of Education and the American Academy of Arts & Sciences.

**Marianne Perie** is the president of Measurement in Practice, LLC, a small education consulting firm focusing on K–12 assessment and accountability. She currently serves on nine state technical advisory committees and the psychometric oversight committee for the AICPA. As an extension of the advisory work, she has provided testimony to state legislatures and boards of education, evaluated standard-setting workshops, facilitated task force meetings, and provided professional development on formative evaluation practices and data literacy. She has consulted with the Council for Chief State School Officers, coordinating the state collaborative on Technical Issues in Large Scale Assessment, serving as a critical friend to states planning new accountability systems, and providing professional development on various assessment issues. Previously, she was the director of two educational research centers at the University of Kansas (KU), overseeing two state operational assessment programs, one career pathway assessment, and several grants, including a research project on the use of learning maps in both assessment and instruction. As a member of KU's graduate faculty in the School of Education, she taught courses in advanced measurement theory and scaling. Prior to joining KU, she was a senior associate with the National Center for the Improvement of Educational Assessment, providing technical assistance to 16 states on accountability and assessment issues related to federal policy. In her early career, she worked on multiple state and district assessments, the National Assessment of Educational Progress, and international assessments as an employee of the Educational Testing Service and the American Institutes for Research.

**Han-Hui Por** is a psychometrician at the Educational Testing Service (ETS) in Princeton, New Jersey, where she tackles psychometric issues in large-scale assessments. Her research interests focus on fairness issues in assessments. Her recent projects include investigating the reliability of human raters' scoring, examining the psychometric properties of new item types and exploring issues of fairness and equity in assessments using process data. She also provides statistical oversight to improve automated scoring systems and new-generation assessments. She started her career in educational assessment with the Ministry of Education in Singapore, where she evaluated the feasibility of investing in new and existing education programs to meet the needs of students, teachers, and the community. Prior to ETS, she was the lead statistician for a New York City Housing Authority study assessing the impact of housing assistance on economic self-sufficiency in New York City. She has also co-authored several papers on the public's understanding and communication of uncertainty and her work has been highlighted by the InterAcademy Council in a report to the United Nations. Por earned her Ph.D. in psychometrics and quantitative psychology from Fordham University, an M.Sc. in applied measurements from the University of Illinois at Urbana-Champaign, and holds a bachelor's degree in psychology and economics from the National University of Singapore.

**Diana C. Pullin** is a research professor and professor, emerita, of education law and public policy in the Lynch School of Education and Human Development and the School of Law at Boston College. The focus of Pullin's work has been the improvement

of access to meaningful educational opportunity for all students. Her research focuses on the impact of law on education practice and the relationship between social science research and professional standards on the law. In addition to her faculty role, Pullin has served as the dean of Boston College's School of Education and as legal counsel for students, educators, and schools in many different types of legal disputes, particularly over high stakes use of testing. She also works as an advisor to lawyers concerning the use of experts in educational and employment litigation involving testing. She has published numerous books, chapters, and articles on education law and public policy, educational and employment testing, educator quality, educational accountability, and individuals with disabilities. She has served extensively as a volunteer expert at the National Academies of Sciences, Engineering, and Medicine advising Congress and state and federal government officials on education policy issues. Pullin is a member of the National Academy of Education, a fellow of the American Educational Research Association, a lifetime national associate member of the National Academy of Sciences, and served on the National Academies' Board on Testing and Assessment. She was formerly the associate editor and co-editor of the interdisciplinary journal *Educational Policy*.

**Stephen G. Sireci** is a distinguished university professor and the director of the Center for Educational Assessment in the College of Education at the University of Massachusetts Amherst. He is currently the president of the National Council on Measurement in Education. He earned his Ph.D. in psychometrics from Fordham University and his master's and bachelor's degrees in psychology from Loyola College Maryland. Before the University of Massachusetts, he was a senior psychometrician at the GED Testing Service, a psychometrician for the Uniform CPA Exam, and a research supervisor of testing for the Newark New Jersey Board of Education. He is known for his research in test construction and evaluation; particularly issues related to content validity, test bias, cross-lingual assessment, standard setting, and computerized-adaptive testing. He has authored/co-authored more than 130 publications, and is the co-architect of the multistage-adaptive Massachusetts Adult Proficiency Tests. He is a fellow of the American Educational Research Association and a fellow of Division 5 of the American Psychological Association. Formerly, he was the president of the Northeastern Educational Research Association, the co-editor of the *International Journal of Testing*, and a senior scientist for the Gallup Organization. He has received several awards, including the Chancellor's Medal (the highest faculty honor at the University of Massachusetts) and the Samuel J. Messick Memorial Lecture Award from Educational Testing Service and the International Language Testing Association. He reviews articles for more than a dozen professional journals and is on the editorial boards of *Applied Measurement in Education*, *Educational Assessment*, *Educational and Psychological Measurement*, the *Journal of Educational Measurement*, and *Psicothema*.

**Marshall S. Smith** is an American educator. He has held academic positions at Harvard University, the University of Wisconsin–Madison, and Stanford University, where he was the dean of the School of Education. He has also held positions in the Ford, Carter, Clinton, and Obama administrations, where he was the Under Secretary and the acting Deputy Secretary.

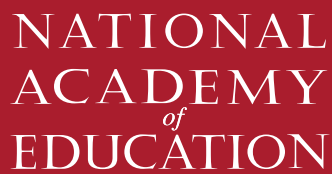
**Jim Soland** is an assistant professor of quantitative methodology at the University of Virginia (UVA) and an associated research fellow at the Northwest Evaluation Association (NWEA), an assessment nonprofit. His work focuses on measurement, growth modeling, and connections between the two, including implications for practice and policy. Applied areas of interest include measuring social emotional learning, understanding how psychological constructs co-develop, and quantifying/correcting for test disengagement. His work has been featured by the Collaborative for Academic, Social, and Emotional Learning and the Brookings Institution. Prior to joining UVA and NWEA, he completed a doctorate in educational psychology at Stanford University with a concentration in measurement and policy. Soland has also served as a classroom teacher, a policy analyst at the RAND Corporation, and a senior fiscal analyst at the Legislative Analyst's Office, a nonpartisan organization that provides policy analysis to support the California Legislature.

**Guadalupe Valdés** is the Bonnie Katz Tenenbaum Professor of Education at Stanford University. Much of her work has focused on the English-Spanish bilingualism of Latinos in the United States and on discovering and describing how two languages are developed, used, and maintained by individuals who become bilingual in immigrant communities. Her books include *Bilingualism and Testing: A Special Case of Bias* (Valdés & Figueroa, Ablex, 1994), *Con Respeto: Bridging the Distance Between Culturally Diverse Families and Schools* (Teachers College Press, 1996), *Learning and Not Learning English* (Teachers College Press, 2001), *Expanding Definitions of Giftedness: Young Interpreters of Immigrant Background* (Lawrence Erlbaum, 2003), *Developing Minority Language Resources: The Case of Spanish in California* (Valdés, Fishman, Chavez, & Perez, Multilingual Matters, 2006), and *Latino Children Learning English: Steps in the Journey* (Valdés, Capitelli, & Alvarez, Teachers College Press, 2010). Valdés is a member of the American Academy of Education and a fellow of the American Educational Research Association. She serves on the editorial boards of a number of journals, including *Modern Language Journal*, *Critical Inquiry in Language Studies*, and *Research on the Teaching of English*.

**Mark Wilson** is a professor of education at the University of California, Berkeley, and also at the University of Melbourne. He received his Ph.D. from the University of Chicago in 1984. His interests focus on measurement. He has published 149 refereed articles, 70 invited chapters in edited books, and 14 books. He was elected the president of the Psychometric Society, and, more recently, the president of the National Council for Measurement in Education. He is also a member of the National Academy of Education, a fellow of the American Educational Research Association, and the American Psychological Association and a national associate of the National Academies of Sciences, Engineering, and Medicine. He is the director of the Berkeley Evaluation and Assessment Research (BEAR) Center. His research program is focused on four mutually supportive areas: (1) the development of a framework for statistical modeling (Explanatory Item Response Modeling) that allows for the extension and adaptation of psychometric models in ways that make them more responsive to problems that arise in education and other areas of application; (2) the application of measurement framework—the BEAR Assessment System (BAS)—to a range of assessment situations across the social sciences, but concentrating on the development of a body of work that can

support the use of sound assessment approaches by teachers and other professionals; (3) the exploration of philosophical and historical perspectives on the area of psychometrics; and (4) the development and dissemination of policy positions in educational testing and assessment, based on the research mentioned above.

**Richard Wolfe** is an associate professor, emeritus, at the Ontario Institute for Studies in Education of the University of Toronto, Department of Applied Psychology and Human Development, Program in Developmental Psychology and Education, in Canada. He studied education measurement, evaluation, and statistical analysis at the University of Chicago. He specializes in planning and evaluating large-scale assessment designs, sampling, data management and analysis, and the use of statistics to improve educational quality. He has served as an advisor and a consultant on testing, methodology, sampling, and psychometrics for government departments and ministries, foundations, nongovernmental organizations, and research institutions in Australia, Canada, Chile, the Dominican Republic, Mexico, Nicaragua, Paraguay, Peru, the United States, Uruguay, and Venezuela. He was the consultant in methodology to the International Mathematics Committee of the IEA Second International Study of Mathematics, an archivist and an analyst of the IEA Second International Study of Science, and the initial technical committee chair of the IEA Third International Study of Mathematics and Science. He worked closely with the TIMSS-associated Study of Mathematics and Science Opportunities. His publications include various papers and books derived from these international studies. He has advised the PISA directorate at OECD on PISA for Schools, school feedback, proposal evaluation, and the mathematics framework. He is a member of the technical advisory group for the California State Department of Education, an advisor and a psychometric analyst for the Ontario Educational Quality and Accountability Office, and has been a member of the expert panel for the U.S. National Assessment Quality Assurance Project.



The **National Academy of Education** (NAEd) advances high-quality research to improve education policy and practice. Founded in 1965, the NAEd consists of U.S. members and international associates who are elected on the basis of scholarship related to education. The Academy undertakes research studies to address pressing educational issues and administers professional development fellowship programs to enhance the preparation of the next generation of education scholars.

[www.naeducation.org](http://www.naeducation.org)